SYSTEMATIC REVIEWS OF RANDOMIZED CONTROLLED TRIALS IN LITERACY RESEARCH: METHODOLOGICAL CHALLENGES

A portfolio presented to the School of Education, University of Sheffield, in partial fulfilment of the requirements for the degree of Doctor of Education (Ed.D.)

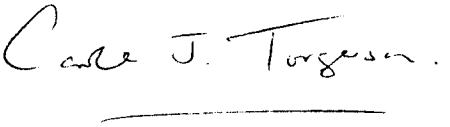
By Carole Joan Torgerson, B.A. (York) M.Litt. (Stirling)

September 2005

Declaration

The research presented in this portfolio was written by the candidate, Carole Joan Torgerson, and has not been submitted in any previous application for a degree.

The portfolio is a record of work undertaken by the candidate. All quotations have been distinguished by quotation marks and all sources of information have been acknowledged.



Carole Joan Torgerson

08-09-2005

Acknowledgements

I am very grateful to my doctoral supervisor Professor Greg Brooks for his guidance and support in the preparation of this portfolio.

The systematic review of phonics instruction (described in Item 3, Chapter 4) was partly funded by the Department for Education and Skills. I undertook all the analyses for the replication, update and secondary analysis of the Ehri *et al* (2001b)[1] phonics review. I developed the original idea and wrote the protocol; I undertook the screening and the data extraction of all the included randomized controlled trials (RCTs). However, I thank and acknowledge: Kath Wright for helping to write the searches for the update and for undertaking them; Jill Hall and Allison Freeman for undertaking double screening and double data extraction for purposes of quality assurance; and Greg Brooks for helping to develop the idea and write the proposal, and for undertaking double data extraction.

I am very grateful to Alison Robinson for proof reading the entire portfolio.

Finally, I am extremely grateful for the support of my family in encouraging me to undertake the Ed. D., and helping me along the way: my parents; my husband, David; and my beloved son, William.

[1]Ehri, L.C., Nunes, S.R., Stahl, S.A. and Willows, D.M. (2001) Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research,* 71(3): 393-447.

Contents

Summary

This portfolio explores the significance of systematic reviews of randomized controlled trials in literacy research. Two methodological challenges to their validity (publication and design bias) form the basis for methodological work, which provides a unique contribution to the educational literature.

Commentary 1 outlines the hypothesis, research questions and significance of the research.

Item 1 has been published as *Systematic Reviews*: a book in the Continuum 'Research Methods' series. It outlines the context of the research: the background to evidence-based education and the rationale for systematic reviews. It gives detailed analyses of the methods for undertaking systematic reviews of randomized controlled trials, with illustrative examples from literacy research undertaken by the author.

Item 2 is a tertiary review which investigates the substantive findings of 14 systematic reviews of effectiveness in literacy learning in English, in relation to their quality. A number of reviews in this tertiary review are judged to be of sufficiently high quality to provide reliable evidence for the effectiveness of literacy interventions and the findings from these reviews are listed in Table 4.

Item 3 assesses the extent to which publication bias and design bias may have affected systematic reviews in literacy learning, using the systematically assembled dataset of reviews from the preceding item. A range of methods is applied to the sample of reviews to examine whether they are susceptible to publication bias. These include the use of funnel and normal quantile plots, and a case study replication, update and secondary analysis of one review to search for any relevant unpublished literature that may have been excluded from the original review. The methodological reliability of individual randomized trials in a sub-sample of the systematic reviews is assessed, and an exploratory meta-regression is undertaken to explore any relationship between study quality and effect size.

Commentary 2 outlines the conclusions and recommendations from the entire body of work. The key messages from the portfolio are: quality appraisal of systematic reviews using a checklist like the QUORUM guidelines may not reveal significant weaknesses in a review. For systematic reviews that have major educational implications, a replication of the review by a second independent research team is warranted, and should be considered as the 'gold standard' for quality appraisal of a review. As a minimum, high quality systematic reviews should routinely report the following: detailed search strategy; search for grey literature; effect sizes and means, standard deviations of means and numbers in each condition for each included trial to enable recalculation of effect sizes. Examination of systematic reviews for publication bias should be undertaken routinely. However, the capacity to detect publication bias is limited when meta-analyses contain small numbers of trials; and interpretation of funnel plots is difficult. Both of these factors should be acknowledged as limitations. Systematic reviews need to be perceived as an objective guide to decision making. However, the possibility of meta-analysis being misleading should always be considered, given the possibility of the introduction of bias in the process of including the studies.

List of Tables

List of Figures

'The house of social science research is sadly dilapidated. It is strewn among the scree of a hundred journals and lies about in the unsightly rubble of a million dissertations. Even if it cannot be built into a science, the rubble ought to be sifted and culled for whatever consistency there is in it.'
Glass *et al* (1981)[2]

'The science of research synthesis – as in any other scientific research - implies that those who practise it will take steps to avoid misleading themselves and others by ignoring biases and the effects of chance.'
Chalmers (2003)[3]

[2] Glass, G.V., McGaw, B. and Smith, M.L. *Meta Analysis in Social Research*, Beverley Hills, CA US, Sage, 1981.
[3] Chalmers, I (2003) Trying to do more good than harm in policy and practice: The role of rigorous, transparent, up-to-date evaluations, *The Annals of the American Academy of Political and Social Science*, 589: 22-40.

Glossary[4]

Intra-cluster correlation (ICC): the statistical correlation between members of the same group (e.g., pupils in the same class).

Meta-analysis: fixed effects model: The fixed effects model of meta-analysis assumes that the variability is exclusively because of random sampling variation around a fixed effect.

Meta-analysis: random effects model: The random effects model of meta-analysis assumes a different underlying effect for each study, and takes this into consideration.

Standard error: The error associated with the estimate of a sample mean. It is calculated by dividing the standard deviation by the square root of the sample size.

Type I error (alpha error): Concluding there is an effect when there is not.

Type II error (beta error): Concluding there is no effect when there is.

[4] A detailed glossary is given in Item 1 of the portfolio (pp. vii – x).

Commentary One:

Research questions; significance of the research; structure and coherence of the portfolio

Contents

Introduction

This portfolio conforms to the University of Sheffield regulations for the submission of a portfolio in partial fulfilment of the requirements for the degree of Ed. D., and comprises the following three linked items:

- Item 1 is a published book (of approximately 21000 words in length).
- Item 2 follows the traditional structure of a journal article (of around 5000 words in length).
- Item 3 is the substantial item (of at least 25000 words in length).

The portfolio critically explores the methodology of systematic reviews of randomized controlled trials (RCTs) as applied to literacy research. The RCT is the most rigorous method for establishing causal relationships. However, in education rarely can one single study definitively answer a research question. This is because of sampling error (random error) due to the very small sample size of so many educational trials. A research synthesis of all the available single studies in a field (which may yield conflicting results) is more likely to be able to come to an overall conclusion about the research effort in a field and detect a causal relationship between independent and dependent variables. Therefore, a systematic review of RCTs is the most appropriate method of addressing the 'what works?' policy question in education.

Although the systematic review has long been established within educational research[5], arguably it has not been as fully utilised, particularly among UK researchers, as it could,

[5] see the list of early reviews in the field of education cited in Glass *et al* (1981), for example: review of teaching style and pupil achievement (Glass *et al,* 1977); review of mainstreaming special education students (Carlberg, 1979) etc., p.23.

or should, have been. Many researchers are equivocal about the use of RCTs within educational research. Consequently, methodological initiatives both within systematic review and trial design have often been driven by other social scientists in recent times, especially healthcare researchers.

This portfolio of methodological work around systematic reviews of RCTs in literacy research aims to transfer methodological developments in systematic review methods that have been developed by other social sciences, particularly in the fields of psychology and health, to systematic reviews in literacy research.

In summary, the entirety of the portfolio aims to strengthen the methodological basis for reviews of RCTs in this field. This commentary describes the research questions, the significance of the research, and the structure and coherence of the portfolio.

Research questions

Systematic reviews of randomized controlled trials in education are essential for distilling and synthesising the research in this field, in order to inform policy, practice and scholarship; but such reviews, though valuable, are susceptible to a range of biases, which may threaten their validity. Such biases need to be assessed and acknowledged. It is the aim of this portfolio to ascertain whether such biases exist in the field of literacy learning. Three research questions arise from this aim. Firstly, are systematic reviews of randomized controlled trials in literacy research of high quality? Secondly, are the randomized controlled trials included in systematic reviews in literacy research of

sufficiently high quality to fit them for purpose? Finally, are the indicators of the quality of randomized controlled trials associated with measures of effectiveness?

Positionality

The positionality of the candidate is set out in detail in Item 1 of the portfolio, particularly in Chapters 1 and 2. In summary, I believe that educational policy and practice should be evidence-based. For narrow issues of what policies, practices, methods, strategies and pedagogies are effective in education the highest form of evidence comes from randomized controlled trials (RCTs). This is because this method is the sharpest tool for probing causal inferences. In questions of effectiveness we are interested in investigating the causal links between intervention and outcome. Random assignment of pupils, students or classes of students is the best method for ensuring that (except for chance differences) comparison groups are equivalent at baseline, and, therefore, any changes observed in the outcome measures can be ascribed to the intervention under investigation.

Following on from Glass *et al* (1981) and Chalmers (2003), I also believe that the single studies in a field should be positioned within all of the available research in a field using the scientific approach of systematic review. Single studies may give misleading results, by chance or due to inadequate methodological design, and therefore, where possible, policy and practice should be based on the results of a synthesis or meta-analysis of several studies.

Significance of the research

The portfolio contains a number of unique methodological advances. For the first time, a quality review of all the meta-analyses in literacy research, and with specific inclusion criteria, has been undertaken. The portfolio also uniquely examines whether or not there is evidence for possible publication bias within trials in literacy research. It also undertakes a replication, update and secondary analysis of an important literacy review, which has not previously been done. Finally, it looks at the association between quality of trials in literacy learning and their effect sizes. This again is a unique contribution to the field.

Structure and coherence: How the three portfolio items are linked and the internal structure of each item

Item 1 has been published as *Systematic Reviews*: a book in the Continuum 'Research Methods' series. The book introduces and develops the rationale for the concepts of both randomized controlled trials and systematic reviews, and justifies the use of systematic reviews of RCTs in educational research. The effectiveness of educational interventions, it is argued, depends upon the systematic synthesis of studies based on the 'gold standard' experimental design – the randomized controlled trial, which is the only design able to create unbiased comparison groups. Epistemological and statistical theories underpin the causal concept that lies at the heart of the RCT. In its ideal form it can be expected to yield an estimate of the 'true' mean effect on a dependent variable. This estimate will deviate from the 'true' effect only by random error, and this should be small when sample sizes are adequate. In this item it is also argued that traditional,

non-systematic reviews may suffer from substantial reviewer bias, which, in turn, may make their results less reliable than systematic reviews. The item describes both the unique systematic, objective features of research synthesis and the practical details in undertaking a systematic review, and it applies quality appraisal criteria adapted from healthcare research to educational research trials. This item also contains a detailed glossary.

In Item 2 the concept of a 'tertiary review' is introduced. This applies systematic review methodology to identifying and summarising existing systematic reviews in literacy learning. The item shows that, by using many of the same methods outlined in Item 1, an overview of all the trial-based research in literacy learning can be undertaken. This type of review is particularly valuable for policy makers who require answers to the broadest questions relating to literacy research. Rather than narrow questions, such as 'Does systematic phonics instruction improve reading?', the tertiary review addresses the broader question 'What interventions are effective in improving literacy learning?'. This tertiary review also produces 'primary data' for the application of detailed methodological work on individual systematic reviews and trials in literacy research in Item 3. The structure of this item follows the traditional structure of a journal article.

In Item 3 methodological innovations are applied to the substantive area of literacy research, often for the first time. In this item the threats to the validity of systematic reviews are explicitly recognised. Unlike in narrative reviews, however, such threats need to be acknowledged and addressed in order to produce a sober interpretation of the literature. A key aim of this item, therefore, is to apply methods that can identify and, sometimes, correct threats to validity.

A major area of methodological work in this item relates to the area of publication bias. Identification of potential publication bias is crucial for responsible interpretation of the primary research literature. The sample of systematic reviews is examined for publication bias using funnel and normal quantile plots, as well as by comparing published and unpublished studies. Quality appraisal criteria developed for healthcare reviews are adapted and applied, for the first time, to reviews in literacy research. The process identifies a review on phonics teaching that appears to suffer from possible publication bias. Given that this area has enormous policy interest both in the UK and the USA, a replication and update of the systematic review is undertaken. In addition to possible publication bias, the replication finds a number of problems with the review, not previously identified by applying quality criteria to the published review. Replication of the review, unexpectedly, identifies several major methodological weaknesses that undermine the conclusions of the original review. This methodological work leads to a number of conclusions and recommendations for further work.

One conclusion from the whole body of work within the portfolio is that for key policy questions review replication is crucial. Systematic review methodology enables replication to be undertaken relatively easily. This important aspect of systematic reviews, it is argued, has not been taken advantage of sufficiently in the past. For example, whilst there have been criticisms of the phonics review by others, such critics have not sought to replicate searching, data extraction and synthesis. Rather they have mainly relied on data contained within the original review. A key finding of this portfolio of work is that this approach, which was adopted in Item 2, may be insufficient, particularly if the question of is of crucial importance. Further research is required to substantiate this finding with other systematic reviews.

The portfolio has contributed to the methodological development of review methods in literacy research in other ways. It is recommended that both funnel and normal quantile plots are used to check for publication bias and, in the case of normal quantile plots, to also check for study heterogeneity.

The basis for any unbiased review is the inclusion of high quality (unbiased) trials. In Item 3 a sample of 56 randomized trials, identified from the systematic reviews from Item 2, is used to assess trial quality in literacy learning. The item finds that many trials are of poor quality, or do not describe their methods in sufficient detail to allow critical methodological scrutiny. An exploratory meta-regression does not find strong evidence of an association between effect size and cluster randomization or attrition. However, the meta-regression finds that there was a reasonably large difference in effect size between studies that used blinded follow-up and those that did not state whether or not follow-up was blind. Recommendations are made on how trials should be conducted and reported so that systematic reviewers are able to access better quality data in the future.

In summary, this portfolio outlines in detail the positionality of the candidate and the methodology of systematic reviewing, and describes the methods for undertaking a systematic review (Item 1). It applies these methods in undertaking a specialised kind of systematic review: a tertiary review (Item 2). Finally, in the last item it applies a range of methodological techniques to reviews and trials in literacy research, and makes recommendations for systematic review methods, trial design and reporting, and further research (Item 3).

Although conclusions and recommendations are described at the end of each item and at the end of each empirical chapter within Item 3, the second commentary sums up all the findings from Items 1, 2 and 3. This item also generates conclusions and makes recommendations for future research into the design of trials and systematic reviews in education.

References

Glass, G.V., McGaw, B. and Smith, M.L. (1981) *Meta-Analysis in Social Research*, Beverly Hills, CA: Sage.

Chalmers, I (2003) Trying to do more good than harm in policy and practice: The role of rigorous, transparent, up-to-date evaluations, *The Annals of the American Academy of Political and Social Science*, 589: 22-40.

Item 1:

Systematic Reviews

Carole Torgerson, Continuum Books, London,

2003

This item is included separately as a published book.

Item 2:

The quality of systematic reviews of effectiveness in literacy learning in English: A 'tertiary' review[6]

[6] A slightly revised version of this item has been submitted to the *Journal of Research in Reading*

Contents

Abstract

Introduction: In this item a 'tertiary' review of systematic reviews in literacy learning is presented. It explores the methodological quality of the identified systematic reviews and identifies the primary data that are used for the in-depth methodological work in Item 3 on the two main threats to the validity of systematic reviews: publication bias and design bias.

Background: Recent governments in the UK have introduced a number of initiatives aimed at improving the literacy levels of children. It is important, therefore, that policy and practice are informed by the most rigorous available evidence, particularly for questions of effectiveness in literacy learning. It is also important that this evidence is subjected to rigorous critical scrutiny.

Methods: Systematic reviews undertaken in the field of literacy learning in English in the years between 1983 and 2003 were searched for, located and quality assessed. The scope of the review was limited to systematic reviews of experimental research evaluating literacy interventions with quantifiable literacy outcome measures in English as a first (not second or additional) language and focusing on children and young people in school settings up to the age of 18.

Results: A total of 14 systematic reviews containing meta-analyses and meeting all the inclusion criteria were included in the tertiary review. The following data were extracted from the reviews: literacy interventions, outcomes evaluated and effect sizes. The quality of the reviews was examined using an adaptation of the QUORUM statement. Overall the quality of the meta-analyses included in this tertiary review was good. When examining the effect sizes of randomized controlled trials (RCTs) and controlled trials (CTs) separately there was no clear pattern as to whether the RCTs produced a larger or smaller effect size than the CTs.

Discussion: Overall the quality of the meta-analyses included in this tertiary review was good. The QUORUM checklist seemed to perform well for the appraisal of educational meta-analyses. All the reviews clearly stated their research question, and their methods of searching for and selecting included studies. Most studies described their data extraction and used some form of quality assessment of included studies. On the other hand, some reviews did have notable methodological weaknesses. Six of the 14 studies did not make an assessment of publication bias, which is potentially a major threat to the validity of any systematic review. In addition, six studies did not provide evidence for reviewer agreement when synthesising the data. There is, therefore, some room for improvement in the methodological quality of systematic reviews in literacy learning.

Conclusions: A number of reviews in this tertiary review are judged to be of sufficiently high quality to provide reliable evidence for the effectiveness of literacy interventions.

Introduction

In this item a 'tertiary' review of systematic reviews in literacy learning is presented. A tertiary review is a systematic review where the included studies are themselves systematic reviews. The tertiary review fulfils two purposes. Firstly, it presents the substantive findings of the identified systematic reviews and explores their methodological quality. Secondly, it identifies the primary data that are used for the in-depth methodological work in Item 3 on the two main threats to the validity of systematic reviews: publication bias and design bias.

Background

Literacy

Recent governments in the UK have introduced a number of initiatives aimed at improving the literacy levels of children. The purpose of literacy instruction is axiomatic: to facilitate the acquisition and development of children's skills in the written language system. In this review, a narrow psychological definition of literacy has been adopted: word identification and recognition, and text comprehension and production, i.e. reading and writing traditional English orthography. It is important, therefore, that policy and practice are informed by the most rigorous available evidence, particularly for questions of effectiveness in literacy learning. It is also important that this evidence is subjected to rigorous critical scrutiny.

Systematic reviews

Systematic reviews are increasingly being seen as important tools for synthesising empirical research, and for influencing policy, practice and future research. The main aim of a systematic review is to gather together in a replicable fashion either all the available evidence on a given subject or a representative sample of the evidence. This evidence may be then combined in some form of synthesis, such as a meta-analysis, in order to give a precise overview of the existing literature within an area. Systematic reviews differ in their methodological rigour. Clearly, a poorly executed systematic review will be less reliable compared with a high quality review. A detailed analysis of the rationale for undertaking systematic reviews in the field of effectiveness research in education is given in Item 1 (Torgerson, 2003), together with a description of their design and conduct. For the purposes of this tertiary review the definitions of 'systematic review' and 'meta-analysis' quoted from Chalmers in Item 1 have been used.

Meta-analysis

Prior to 1976 when Glass (1976) first described the procedures for meta-analysis (see also Glass *et al*, 1981) there was no method for resolving the variability in individual studies when combining studies in a review. Glass proposed a statistical method for aggregating the data from individual studies: 'meta-analysis'. In a meta-analysis the results from individual studies with similar conceptual underpinnings but with different measurement scales for outcomes are expressed in a standard metric: an effect size, which is the size of a treatment effect. The effect size is usually the difference between the means of the experimental and control groups at post-test divided by the standard deviation of the control group (Glass, 1976). Hedges (1989) used the pooled standard

deviation as being 'more stable'. This will be because a pooled standard deviation will give a better estimate of the standard deviation because essentially it will be based on two samples. The effect size (expressed in standard deviation units) illustrates the magnitude and direction of the effect of a treatment (Cohen, 1977). In a meta-analysis all the effect sizes are pooled into an aggregate effect size which weights the individual studies by their sample sizes. In addition, subgroup analyses can be undertaken to explore the effect of moderator variables. One of the strengths of meta-analysis is that it includes a record of judgements made at each stage of the process, including the warrants for the conclusions. Also it can combine results from a number of studies to give an overall interpretation of the research. Meta-analysis is limited by the quality of its included studies (which may be poor), although it does have the benefit of 'correcting' for some of the limitations in the individual trials.

Quality of systematic reviews

Because the robustness of the findings of systematic reviews is underpinned by the quality of their design, conduct and reporting, it is important for any tertiary review to consider the methodological quality of the systematic reviews that it includes, which depends upon how they deal with the methodological variability of their included studies. Important quality questions include, for example, whether or not the reviewers examined the effect sizes by method of allocation (random or other method) and whether or not reviewers examined the results of their findings in relation to the methodological characteristics of the included studies. The evaluation of the quality of systematic reviews includes questions to ascertain the degree to which bias has been limited in the review such as: whether or not the review question is significant and conceptually underpinned by previous empirical or theoretical work; whether the

searching was exhaustive (including a search of 'fugitive' or 'grey' literature in order to limit the possibility of publication bias) and transparent; and whether the variability in the quality of included trials was assessed and taken into consideration in the synthesis.

The importance of the quality appraisal of systematic reviews has been recognised in the field of healthcare research, where a set of guidelines known as the QUORUM (Quality Of Reporting of Meta-analyses) statement has been developed (Moher *et al*, 1999; Shea *et al*, 2001). Like the CONSORT statement (Altman, 1996) for the reporting of RCTs, the QUORUM statement was developed by methodologists as a consensus statement for the quality of reporting of meta-analyses in healthcare research, and has been adopted by a number of mainstream healthcare journals. It is believed that the quality of reporting of meta-analyses is a reasonably good (though not perfect) indicator of the quality of the review (Shea *et al*, 2001). In Item 1 (pp. 70-72) the main points of the QUORUM statement are outlined. The stages of a meta-analysis in which the QUORUM standards should be adopted are: the rationale for the meta-analysis; the methods and results for the search; the inclusion/exclusion criteria; the coding of the primary studies; the meta-analysis; quality assurance procedures; and the interpretation of the results. The checklist was developed, where possible, from previous empirical research evidence regarding the possibility of a biased meta-analysis resulting from the failure to report any of the areas highlighted. In the revised QUORUM statement, researchers, editors and peer referees are provided with a list of 18 items to consider in the reporting of meta-analyses under the broad headings: title, abstract, introduction, methods, results and discussion.

Other quality appraisal checklists are available, particularly in the area of healthcare research. For example, a systematic review on the reporting of RCTs in healthcare research found thirty-four checklists (Shea *et al*, 2001). More recently, the Campbell Collaboration has developed guidelines for the writing of protocols for systematic reviews in the fields of education, criminal justice and other social sciences; and in the UK the EPPI-Centre has developed similar guidelines specifically in the field of education. However, for this item an adaptation of the QUORUM checklist has been used (Table 1). This checklist is specifically designed for the quality appraisal of meta-analyses of experimental research and is therefore more appropriate for its purpose in this item. The QUORUM statement has been modified in the following ways. The checklist adapted for educational meta-analyses is organised under four broad headings: introduction, methods, results and discussion, and includes all 11 items in the QUORUM statement under these headings. In order to make the guidelines relevant to educational research, educationally appropriate terminology has been adopted to describe participants and interventions.

Table 1: Key features of the QUORUM statement (adapted for systematic reviews in education)

Introduction: Explicitly state educational problem and rationale for review.
Methods: Searching: State sources of information (e.g., names of databases, hand searching of key journals), search restrictions (e.g., year, publication language, published and/or unpublished).
Selection: Inclusion and exclusion criteria.
Validity assessment: Quality assessment (e.g., blinded follow-up).
Data abstraction: Process used (e.g., double data extraction).
Study characteristics: Type of study design, student characteristics, details of intervention, outcomes, how was educational heterogeneity assessed?
Data synthesis: How were data combined? Measures of effect, statistical testing and confidence intervals, handling of missing data, sensitivity and subgroup analyses, assessment of publication bias.
Results: trial flow: Provide a profile of trials identified and reasons for inclusion/exclusion.
Study characteristics: Provide descriptive data for each trial (e.g., age, setting, class size, intervention).
Quantitative data synthesis: Report agreement between reviewers on selection and validity assessment, present summary results, report data needed to calculate effect sizes and confidence intervals (i.e., number, mean, standard deviations by group).
Discussion: Summarize key findings and educational inferences. Interpret results in light of all the evidence, acknowledge potential biases in review and suggest areas for future research.

Source: adapted from Shea *et al* (2001)

Effectiveness research

The most reliable method of establishing a causal connection between a teaching intervention in literacy learning and literacy outcomes is some form of experimental research. In Item 1, the use of systematic reviews of randomized controlled trials (RCTs) is justified as being the 'gold-standard' approach in questions that assume a connection (Torgerson, 2003; Cook and Campbell, 1979; Shadish *et al*, 2002). In the following, the principal arguments from Item 1 are briefly rehearsed.

In the wider research community, research questions focusing on effectiveness use the randomized controlled trial (RCT), which is the strongest experimental design. A 'gold-standard' review is a systematic review of all the randomized controlled trials (RCTs) in a field. To produce evidence with respect to literacy learning requires either a systematic review of all the RCTs in the field or, alternatively, a tertiary review of existing systematic reviews. The latter is likely to be more efficient in terms of time and resources and could enable a policy maker to survey the whole of evidence on a disparate range of interventions for different outcomes in literacy learning. When two

or more RCTs that are sufficiently homogeneous are identified in a systematic review they are usually combined in a meta-analysis. Combining studies in a meta-analysis increases the precision of the estimate of the effect over any estimate from a single study. Also, having more than one experimental replication of an intervention will reassure a reader that the intervention can be transferred to another location and time. In this item, therefore, only systematic reviews that undertake a meta-analysis will be included.

In summary, the aims of this item are: to identify all the relevant systematic reviews in literacy learning; to report their substantive findings and quality appraise them using the adapted QUORUM statement; and to assemble a systematically retrieved dataset of reviews in literacy learning for methodological work in Item 3 of the portfolio.

Methods

In this item systematic review methods have been used throughout as outlined in Torgerson (2003), Item 1 of the portfolio, in order to limit bias. As with all systematic reviews, the conceptual underpinning for the review and its methods are explicitly described according to a replicable methodology.

Systematic reviews undertaken in the field of literacy learning in English in the years between 1983 and 2003 have been systematically searched for, located, and quality assessed.

It is important to make the assumptions underlying the decision to include or exclude a systematic review at each stage of this tertiary review explicit. The decision was made

to focus on the use of randomized controlled trials to compare the relative effectiveness of different literacy interventions by isolating the factors in a causal relationship between intervention and outcome. Therefore, to be eligible for inclusion for this tertiary review, candidate systematic reviews had to include at least one RCT and a pooled effect size (meta-analysis).

The scope of the review has been limited to systematic reviews (written in English) of experimental research evaluating literacy interventions with quantifiable literacy outcome measures in English as a first (not second or additional) language and focusing on children and young people up to the age of 18. Where reviews also included studies where the participants were in college or university settings (rather than school settings) these were only included in the tertiary review if separate effect sizes were presented for the two populations of participants.

Inclusion/exclusion criteria

Systematic reviews were included in the tertiary review if they fulfilled four basic criteria: the reviews had to state both their methods for searching and their inclusion and exclusion criteria; they had to include some form of quality appraisal (however minimal) of the included studies; and they had to quantitatively synthesise the results of the included studies by reporting average effect sizes across the studies. Decisions about whether or not to include systematic reviews at both first and second stages of the review were made independently of the results of the reviews. A summary of the inclusion and exclusion criteria for the tertiary review is given in Appendix A.

Search methods

Three electronic databases were searched at the end of 2003: PsycINFO; The Educational Resources Information Center (ERIC); and The Campbell Collaboration Social, Psychological, Educational Criminological Trials Register (C2 SPECTR). The search included the years between 1983 and 2003. The key words used in the search included: systematic review or meta-analysis; best evidence synthesis or research synthesis; literacy or reading or writing. After all the electronic searches were completed, the bibliographic details and abstracts of all the initial 'hits' were exported from each database and imported into EndNote, where duplicate references were removed.

Potential systematic reviews were identified from the electronic searches in two stages. They were first screened on the basis of their titles and abstracts, using pre-established inclusion/exclusion criteria. Papers that appeared to be relevant were sent for through interlibrary lending and read in full. They were then re-screened on the basis of the full paper, again using the inclusion/exclusion criteria. The bibliographies of previous tertiary reviews and any included systematic reviews were also scanned for potentially relevant reviews that were then sent for and re-screened. Any potentially relevant reviews identified through 'contact' were screened on the basis of the inclusion/exclusion criteria. The cut-off date for receipt of papers was December 31st 2004. This date was established for pragmatic reasons.

Data extraction and quality assessment

Data were extracted from all reviews included after the first and second stages of screening using a pre-established data extraction form which included the following criteria: the nature of the literacy aspect(s) investigated; the number and design of the

included studies; whether or not it was possible for the reviewer to distinguish between the RCTs and the CTs in the paper without sending for all the included experimental studies; setting(s), intervention(s) and participants(s); outcome measures; results (including effect sizes and confidence intervals (CIs), standard errors (SEs) or statements about statistical significance); conclusions. All included systematic reviews were quality assessed using an adaptation of the QUORUM statement (see Table 1).

Results

The searches identified 206 references for possible inclusion in this review. Four further references for possible inclusion were obtained through contact or citation, making a total of 210 to be screened at first stage. Ninety-five studies were included and sent for through inter-library lending. Ten papers were unobtainable or not received by the cut-off date. After screening of the full papers 14 systematic reviews meeting all of the inclusion criteria were included i.e., they were systematic reviews including a quantitative synthesis (meta-analysis) of experiments (including at least one RCT) evaluating literacy interventions in populations of English as a first language learners up to the age of eighteen. The remaining 196 papers were either tertiary reviews, were not received, did not contain a quantitative synthesis, were not systematic reviews of literacy interventions with literacy outcomes, did not contain any RCTs or contained non-experimental research, or they did not focus on learners with specific characteristics as specified in the pre-determined criteria. For example, any quantitative research synthesis of experimental research that did not mention the use of random allocation to generate intervention and control groups as a variable was excluded. Although many of the excluded reviews were high quality integrative reviews containing valuable insights in the field of literacy learning, they did not fulfil the strict criteria that determined

inclusion in this review. A full list of all the 196 studies that were either not received or excluded from the tertiary review is given in Appendix D. The two other tertiary reviews that were located (Guthrie *et al*, 1983; Lipsey and Wilson, 1993) are briefly described in order to distinguish them from this tertiary review. Guthrie *et al* (1983) undertook a narrative synthesis of reviews of reading research, and coded them by content of reviews, publication outlet and citation rate in order to assess the potential influence of reviews by reading topic. Lipsey and Wilson (1993) systematically assembled a body of 302 meta-analyses in psychological, educational and behavioural treatment research over a period of 15 years, and looked for potential sources of bias in the meta-analyses. This tertiary review included six reviews in the field of literacy learning, although none of these reviews fulfilled the criteria for inclusion in my review.

Table 2 shows the characteristics of the included studies. These characteristics include the aspects of literacy studied and the effect sizes as presented in the reviews. As the table shows, there were 10 different interventions included in the 14 reviews: computer-based learning; reading aloud to young children at school; parents reading to pre-school children; phonological awareness training; phonemic awareness training; the use of volunteers in literacy learning; writing instruction; meta-cognitive instruction for reading; whole language reading; and peer tutoring.

To assess whether or not RCTs and CTs were producing different conclusions, where possible separate effect sizes for randomized controlled trials and other controlled trials are presented in the table. Also given are the reviewers' conclusions, where possible as direct quotations from the reviews.

Table 2: Characteristics of the 14 included reviews

Author, date, country	Aspect(s) of literacy (independent variable)	Number and design of studies	Able to distinguish between RCTs and CTs in meta-analysis?	Setting(s)	Outcome measure(s) (dependent variable(s))	Results, with mean weighted effect sizes (statistical significance or CI or SE reported) Effect sizes for CTs and RCTs reported separately if present in report	Conclusions
Bangert-Drowns (1993), USA	Writing: word-processing in writing instruction.	32 reports containing 33 studies: 'true' and quasi-experiments.	N*	Preschool – college (Elementary, Junior High, High, College);	Quality of writing; number of words; attitude towards writing; adherence to writing conventions; frequency of revision.	Quality of writing = (20) 0.27 (statistically sig. p=0.02; SE=0.11); number of words = (5) 0.52 (statistically significant); attitude towards writing (9) = 0.12 (SE = 0.21, not statistically sig.); frequency of revision (4) = 0.18 (SE = 0.36) **Pre-college = 0.36 (SE = 0.20); college = 0.18 (SE = 0.09) **Quality of writing: CT (8 studies) = 0.39 (SE = 0.18, stat. sig.)** **RCT (7 studies) = 0.31 (0.22, not stat. sig.).**	'Word processing groups, especially the weaker writers, improved the quality of their writing. Word processing students wrote longer documents but did not have more positive attitudes towards writing..' (p.69). 'Basic writers tended to benefit more from word processing than higher ability students... basic writers tended to have less variance in writing quality after experience with word processing… duration of this experience did not appear related to writing quality' (p. 87).
Blok (1999), Netherlands	Reading: reading to young children in educational settings.	10 studies containing 11 independent samples – randomized, matched and non-randomized	N	School settings; 31-90 months.	Oral language; Reading.	Oral language = 0.63 (p<.01; SE = 0.14) Reading = 0.41 (p<.01; SE = 0.15)	Reading to young children at school has a moderate effect on their reading development. 'Although these figures look promising, caution is needed because the empirical evidence appears to be meagre. Not only is the number of studies small,

		designs.					but a critical analysis of the design of the studies generally reveals poor quality' (p.343). 'Reading at school has a moderate to large effect on children's oral language development and a moderate effect on their reading development. Both outcomes are statistically significant, because zero does not fall within the 95% reliability interval' (p.363).
Bus *et al* (1995), Netherlands	Reading: parent-preschooler reading.	29 studies: correlational, longitudinal, experimental, retrospective.	**N**	Parent-preschool settings.	Language and reading skills.	Combined studies = 0.59 Book reading and language skills studies = 0.67 Book reading and emergent literacy studies = 0.58 Book reading and reading achievement = 0.55.	'This study shows that book reading is effective' (p.17). Results give 'straightforward support for family literacy programs' (p.15). 'Parent-preschool reading is related to outcome measures such as language growth, emergent literacy and reading achievement…' 'There are hardly any studies with negative effects, indicating that book reading has a positive effect on outcome measures' (p.15). In addition the effect was not dependent on socio-economic variables.
Bus and van IJzendoorn (1999), Netherlands	Reading: phono-logical awareness training.	34 studies (15 RCTs and 19 CTs).	**Y**	School settings (kindergarten and primary schools).	Phonological awareness; reading.	**'In a homogeneous set of US studies with a randomized or matched design, the combined effect sizes for phonological awareness**	Phonological awareness is an important but not a sufficient condition for early reading.

						and reading were *d* = 0.73 (stat. sig.) and *d* = 0.70 (stat. sig.) respectively.' (p.403).	
Ehri *et al* (2001a), USA	Reading: phonemic awareness instruction.	52 studies.	**N***	Pre-school to 6th grade.	Phonemic awareness; reading; spelling	The effect of phonemic awareness instruction on helping children to acquire phonemic awareness was large and statistically significant (d = 0.86); phonemic awareness instruction had moderate, statistically significant effects on reading (d = 0.53) and spelling (d = 0.59). **PA outcomes: random assignment: d = 0.87; matched design d = 0.92, non-equivalent design d = 0.83. Reading outcomes: random assignment d = 0.63, matched design d = 0.57, non-equivalent design d = 0.40. Spelling outcomes: random assignment d = 0.37, matched design d = 0.73, non-equivalent design 0.86.**	'Phonemic awareness instruction was found to make a statistically significant contribution to reading acquisition' (p.251). 'Our findings indicate that teaching phonemic awareness is a means rather than an end. PA is taught not for its own sake but rather for its value in helping children understand and use the alphabetic system to read and write. Findings also indicate that a moderate amount of time rather than a huge amount of time may be sufficient to teach PA' (p.279).
Ehri *et al* (2001b), USA	Reading: systematic phonics instruction	38 studies (13 RCTs and 25 CTs)	**Y**	Kindergarten to 6th grade.	Word reading outcomes, measures of reading fluency, comprehension and spelling.	'The overall effect of phonics instruction on reading was moderate d = 0.41' (p.393) **Random assignment d =**	'Findings of the meta-analysis support the conclusion that systematic phonics instruction helps children learn to read more effectively than non-

						0.45, non-equivalent groups d = 0.43.	systematic or no phonics instruction' (p.427).. 'Systematic phonics instruction helped children to read better than all forms of control group instruction, including whole language instruction. In sum, systematic phonics instruction proved effective and should be implemented as part of literacy programs to teach beginning reading as well as to prevent and remediate reading difficulties' (p.393).
Elbaum and Vaughn (2000), USA	Reading: adult-instructed one-to-one reading instruction.	29 studies	**N***	Elementary students at risk for reading failure.	Reading comprehension and other reading skills; spelling; writing.	Reading outcomes had a mean weighted effect size of 0.41 (CI 0.32 to 0.49). **'Studies that used random assignment or matching yielded significantly higher effect sizes (d = 0.56, CI 0.45 to 0.66) than studies that used other procedures (e.g. teacher judgement, convenience: d = 0.17, CI 0.04 to 0.30)' (p.613).**	'The meta-analysis revealed that college students and trained, reliable community volunteers were able to provide significant help to struggling readers. This finding suggests that it may be possible to reduce the cost of providing effective supplemental, one-to-one instruction to students at risk for reading failure.' (p.616).
Gersten and Baker (2001), USA	Writing instruction in expository and narrative text.	13 'true' and quasi-experiments.	**Y**	Students with learning disabilities.	Writing performance.	Mean effect size across all studies: 0.81 (CI 0.65 to 0.97); **RCTs had a weighted effect size of 1.19 (CI 0.83 to 1.55), CTs had a weighted effect size of 0.71 (CI 0.53 to 0.89).**	'Results indicated that the interventions used in the research studies consistently produced strong effects on the quality of students' writing…' (p.251).
Haller *et al*	Reading:	20 studies.	**N**	2nd to 12th grade	Meta-cognitive skills.	Mean effect size across all	Results indicated that 'although

(1988), USA	meta-cognitive instruction in reading comprehen--sion.					studies was 0.71 (SD = 0.81).	meta-cognitive instruction was helpful at all grade levels, it seemed particularly effective for seventh and eighth graders' (p.8). '...several meta-cognitive strategies are particularly effective. These include awareness of textual inconsistency and the use of self-questioning as both a monitoring and a regulating strategy' (p.8).
Jeynes and Littell (2000), USA	Reading: whole language instruction.	14 studies.	N	School settings – kindergarten – G3.	Change in literacy level (reading achievement).	Total effect size for low-socioeconomic status (SES) children receiving whole language versus children receiving basal instruction = -0.65 (-0.61 to -0.69), significant at the 0.0004 level; for standardised tests only = -0.70 (-0.74 to -0.07) p<0.001; for non-standardised tests only = 0.13 (-0.02 to -0.48) p=0.025.	'Overall, the evidence suggested that low-SES primary school children do not benefit from whole language instruction, compared to basal instruction. Nevertheless, the results indicated that there may be some advantages to the whole language approach in its purest form.' (p.21). 'For all the whole language studies combined, low-SES children receiving basal instruction did consistently better on the various literacy measures than their counterparts who received whole language instruction' (p.25-6).
Mathes and Fuchs (1994), USA	Reading: peer tutoring.	11 studies: 9 RCTs; 2 CTs.	Y	School-age students, grades 1-12, students with disabilities in reading.	Reading achievement.	Average effect size: 0.36 (p<0.01); no-research treatment control 0.40, research-treatment control 0.14.	'Peer tutoring was found to have an overall effect size of 0.36 and to be more effective than typical reading instructions regardless of setting. Peer tutoring was not, however,

							more effective when contrasted to other, teacher-led interventions, such as one-to-one teacher tutoring or teacher-led small group instruction,' (p.59).
Torgerson and Elbourne (2002), UK	Spelling: computer-based instruction.	6 studies (all RCTs).	**Y (no CTs).**	1st to 6th grade.	Spelling test.	**Pooled effect size: 0.37 (-0.02 to 0.77, p=0.06).**	'This review suggests that the teaching of spelling by using computer software may be as effective as conventional teaching of spelling, although the possibility of computer-taught spelling being inferior or superior cannot be confidently excluded due to the relatively small sample size of the identified studies' (p.129).
Torgerson *et al* (2002), UK	Reading: volunteer tutoring.	7 studies (all RCTs).	**Y (no CTs).**	1st to 6th grade.	Reading comprehension.	**Pooled effect size: 0.19 (-0.31 to 0.68, p=0.54).**	'Overall, volunteering appeared to have a small effect on reading outcomes. However, the confidence intervals were wide, which could conceal a potentially large benefit or a harmful effect' (p.433).
Torgerson and Zhu (2003), UK	Reading, writing and spelling: computer-based instruction.	16 studies (all RCTs).	**Y (no CTs).**	Pupils and young people aged 5 to 16	Reading; Writing; Spelling.	**Spelling and CAI pooled effect size: 0.20 (-0.18 to 0.58); Word-processing and writing: 0.89 (0.25 to 1.54); Computer-mediated texts and comprehension (2 effect sizes included because outcome measures judged to be equally educationally**	'Teachers should be aware that there is no evidence that non-ICT methods of instruction and non-ICT resources are inferior to the use of ICT to promote literacy learning' (p.9).

		significant, and effect sizes are going in opposite directions): -0.05 (-0.33 to 0.24); 0.28 (-0.003 to 0.57)

Y = Yes
N = No
*N = No, but effect sizes for RCTs and CTs reported separately
** = results reported separately for each level of education
CI = Confidence intervals
SE = Standard error

Table 3: Quality appraisal of 14 included reviews, using an adaptation of the QUORUM statement

Study	Intro.	Methods: search	Methods: selection	Methods: validity assessment	Methods: data abstraction	Methods: study character-istics	Methods: data synthesis	Results: trial flow	Results: study character-istics	Results: data synthesis	Discussion
Bangert-Drowns (1993)	Y	Y	Y	Y	NS	Y	Generally Y; but no assessment of publication bias	NS	Y	Y: summary results; but NS: agreement on selection and validity assessment	Y
Blok (1999)	Y	Y	Y	Y	NS	Y	Y	NS	NS	Y: summary results; but no raw data	Y
Bus *et al* (1995)	Y	Y	Y	NS	NS	Y	Y	NS	Y	Y: summary results; NS: agreement between reviewers on selection and validity assessment	Y
Bus and van IJzendoorn (1999)	Y	Y	Y	Y	Y	Y	Generally Y; but no assessment of publication bias	NS	Y	Y	Y: key findings and implications; NS potential biases
Ehri *et al* (2001a)	Y	Y	Y	Y	Y	Y	Generally Y; but no assessment of publication bias	NS	Y	Y	Y
Ehri *et al*	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

(2001b)											
Elbaum Vaughn (2000)	Y	Y	Y	NS	Y	Y	Generally Y; but no assessment of publication bias	NS	Y	Y	Y
Gersten and Baker (2001)	Y	Y	Y	Y	Y	Y	Y; N: no assessment of publication bias	NS	Y	N: no reporting of inter-rater agreement; no raw data	Y: key findings summarised; N: potential biases in review not acknowledged
Haller *et al* (1988)	Y	Y	Y	Y	NS	Y	Generally Y; but no assessment of publication bias	NS	NS	N: summary data; raw data; inter-rater agreement	Y: key findings; N: potential biases in review not acknowledged
Jeynes and Littell (2000)	Y	Y	Y	Y	Y	Y	Y	NS	NS	Y: summary data; N: raw data	Y
Mathes and Fuchs (1994)	Y	Y	Y	Y	Y	Y	Generally Y; but no assessment of publication bias	NS	Y	Y: summary of key findings and interpretation; N: raw data not presented	N: potential biases in review not acknowledged
Torgerson and Elbourne (2002)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Torgerson *et al* (2002)	Y	Y	Y	Y	Y	Y	Generally Y; but no assessment	Y	Y	Y	Y

							of publication bias				
Torgerson and Zhu (2003)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

Y: Yes
N: No
NS: Not stated

Table 3 presents the quality assessments of the reviews. As the table shows, most of the reviews tended to be of high quality. All of the reviews set clear research questions and described their search strategies and their inclusion and exclusion criteria clearly. Some reviews (6), however, did not examine the possible influence of publication bias on their results. Three reviews also did not acknowledge the role that potential biases might have on their findings and conclusions.

Randomized controlled trials versus quasi-experiments

In seven of the 14 reviews it was not possible to distinguish between the included randomized controlled trials and studies of other designs (including correlational studies and quasi-experiments). In three of these seven reviews, although it was not possible for the reader to distinguish between studies of different types, the reviewers had given separate effect sizes for randomized controlled trials and controlled trials. However, in some cases not all of the included studies were included in this analysis. In the other seven reviews, either it *was* possible for the reader to distinguish between the different study types or the review only contained randomized controlled trials. Table 4 presents summaries of all the reviews where effect sizes are available separately for randomized (or matched) trials and controlled trials. Also included are details about the literacy interventions, the primary outcomes, and the results.

Table 4: Pooled effect sizes according to design of study i.e. whether RCTs, studies with matched group design or CTs.

Author (date)	Literacy intervention	Pooled overall effect size	Pooled effect size of RCTs	Pooled effect size of matched groups design	Pooled effect size of CTs	Summary
Bangert-Drowns (1993)	The effect of word processing on the quality of writing	0.27	0.31	-	0.39	A statistically significant positive effect for word processing on the quality of writing.
Blok (1999)	The effect of reading to young children at school on their reading development	0.41	-	-	-	A statistically significant positive effect for reading to young children on oral language and reading development.
Bus et al (1995)	The effect of parent pre-school reading on reading skills	0.59	-	-	-	A positive effect for book reading on reading and language outcomes.
Bus and van IJzendoorn (1999)	The effect of phonological awareness training on reading	0.70	0.70	-	-	Phonological awareness is an important but not a sufficient condition for early reading.
Ehri et al (2001a)	The effect of phonemic awareness instruction on reading outcomes	0.53	0.63	0.57	0.40	Phonemic awareness instruction was found to make a statistically significant contribution to reading acquisition.
Ehri et al, (2001b)	The effect of systematic phonics instruction on reading outcomes	0.41	0.43	-	0.45	Systematic phonics instruction helps children learn to read more effectively than non-systematic or no phonics instruction.
Elbaum and Vaughn (2000)	The effect of one-to-one reading instruction on reading comprehension	0.41	0.56	-	0.17	Volunteers were able to provide significant help to struggling readers.
Gersten and Baker (2001)	The effect of writing instruction in expository and narrative text	0.81	1.19	-	0.71	Writing instruction consistently produced strong effect on the quality of students' writing.
Haller et al (1988)	The effect of meta-cognitive instruction in reading comprehension.	0.71	-	-	-	Meta-cognitive instruction was helpful at all grade levels, particularly at seventh and eighth grades.
Jeynes and Littell (2000)	The effect of whole language instruction on reading	- 0.65	-	-	-	A statistically significant negative effect for whole language instruction.

	outcomes					
Mathes and Fuchs (1994)	The effect of peer tutoring on reading	0.36	-	-	-	Peer tutoring was found to be more effective than typical reading instruction regardless of setting.
Torgerson and Elbourne (2002)	The effect of computer-based instruction on spelling	0.37	0.37	-	-	The teaching of spelling using computer software may be as effective as conventional teaching, although the possibility of computer-taught spelling being inferior or superior cannot be excluded.
Torgerson et al (2002)	The effect of volunteer tutoring on reading	0.19	0.19	-	-	Overall, volunteering appeared to have a small effect on reading outcomes.
Torgerson and Zhu (2003)	The effect of computer-based literacy instruction	0.20 (spelling) 0.89 (word-processing) -0.05 (computer-mediated texts)	0.20 (spelling) 0.89 (word-processing) -0.05; 0.28 (computer-mediated texts)	-	-	Teachers should be aware that there is little evidence that non-ICT methods of instruction and non-ICT resources are inferior to the use of ICT to promote literacy learning.

Discussion

This methodological tertiary review identified 14 eligible systematic reviews and meta-analyses. These reviews were appraised using the QUORUM checklist, which was developed for systematic reviews in healthcare research. The checklist seemed to perform well for the appraisal of educational meta-analyses.

Study quality

Overall the quality of the meta-analyses included in this tertiary review was good. All the reviews clearly stated their research question, and their methods of searching and selecting included studies. Most studies described their data extraction and used some form of quality assessment of included studies. The generally good quality of all included reviews may be a consequence of the strict inclusion criteria. For example, the need to separately identify randomized and non-randomized trials meant that reviews that did not differentiate between the two types of controlled trial probably had other weaknesses.

On the other hand, some reviews did have notable methodological weaknesses. Six of the 14 studies did not make an assessment of publication bias, which is potentially a major threat to the validity of any systematic review. In addition, six studies did not provide evidence for reviewer agreement when synthesising the data. There is, therefore, some room for improvement in the methodological quality of systematic reviews in literacy learning.

Some reviews included both RCTs and CTs. When examining the effect sizes of solely RCTs and CTs there was no clear pattern as to whether the RCTs produced a larger or

smaller effect size than the CTs. In some reviews (e.g., Gersten and Baker, 2001) meta-analyses of RCTs found a larger effect size than the corresponding analysis of CTs. In contrast, Bangert-Drowns (1993) found smaller effect sizes for RCTs than for CTs.

Substantive findings

Although the main aim of this review was to evaluate the quality of systematic reviews in literacy learning, a secondary aim was to map the substantive findings. A number and variety of interventions were found to be effective in improving literacy development in children and young people. These interventions were evaluated through systematic reviews and meta-analyses of variable quality, although the rigorous methodological criteria for inclusion in the review should be stressed, leading to the conclusion that all of the included reviews are of a reasonably good quality.

ICT and literacy

Three reviews looked at the use of computer technology to improve literacy. One of these reviews contained three separate meta-analyses (Torgerson and Zhu, 2003). Therefore there were a total of five separate meta-analyses in this area. Two (Torgerson and Elbourne, 2002; Torgerson and Zhu, 2003) found little evidence of benefit for the use of ICT in spelling or reading, whilst two found that word processing helped weaker writers to improve the quality of their writing (Bangert-Drowns, 1993; Torgerson and Zhu, 2003). Generally the quality of reporting of meta-analyses in the field of computer-based instruction in literacy learning was high, with all reviews reporting the justification for the review, the methods and results. The only exception was the review by Bangert-Drowns (1993) which did not state the methods for extracting data from the included studies or include a flow diagram of all the included and excluded studies, which makes replication of a review difficult. Also, there was no assessment of

publication bias. However, this review was strong in all other aspects of reporting. In summary, the quality of evidence for the results of meta-analyses in ICT and literacy is high and therefore it is likely that their results are reliable. The use of word-processing for improving the quality of writing seems to be a promising intervention, particularly for weaker writers.

Reading aloud to young children

Two reviews looked at the evidence for the effectiveness on literacy outcomes of interventions involving adults reading to children. One study (Blok, 1999) found some evidence to support the use of reading aloud to children in school settings, although the authors were cautious in this interpretation given their appraisal of the quality of the studies and concluded that there was relatively weak evidence in the area. The second study (Bus *et al*, 1995) found evidence of a benefit of parents reading aloud to their pre-school children. As Table 3 shows, both of these reviews had methodological strengths but lacked detail in some of their reporting. Although both reviews contained randomized controlled trials, and acknowledged these as the highest form of evidence, neither of them reported separate effect sizes for different study designs. However, Bus *et al* (1995) examined their data for a differentially larger pooled effect size in experimental studies but did not find it. Whilst the positive effect on language outcomes of reading to children is large in either setting, the positive effect of parental reading on reading outcomes is larger than the effect of school reading on reading outcomes, a finding that Blok (1999) highlights in his analysis. Finally, although it can be concluded from these two meta-analyses that there is some evidence that reading aloud to young children in both home and school settings is effective in improving language and reading outcomes, it should be noted that Blok (1999) in particular is

cautious about the strength of the evidence base due to methodological weaknesses of some of the included studies.

Phonemic awareness training and phonics instruction

Two reviews looked at phonemic awareness training (Ehri *et al*, 2001a; Bus and van IJzendoorn, 1999) and one examined the effectiveness of systematic phonics instruction (Ehri *et al*, 2001b). Two of these three reviews did not assess possible publication bias or acknowledge the potential for bias in the review (Bus and van IJzendoorn, 1999; Ehri *et al*, 2001a). One of the reviews did not report the flow of included and excluded studies in the review (Ehri *et al*, 2001a). However, apart from this these reviews were rigorous in their reporting of rationale, methods and results. Therefore, on the basis of their reporting, there is good quality evidence that the results and conclusions from these reviews are reliable. They all concluded that there is a benefit of a structured approach to phonemic awareness instruction and systematic phonics instruction in early literacy, particularly in populations of children at risk of reading failure.

Whole language instruction

The one included review that looked at the effectiveness of whole language instruction in reading compared with 'basal instruction' did not find the approach beneficial (Jeynes and Littell, 2000). However, although Jeynes and Littell did not report study characteristics and separate effect sizes for randomized and non-randomized designs, the overall quality of this review was good. The authors outlined some of the limitations to their meta-analysis, including: difficulty in defining 'whole language' interventions; the difficulty with the implementation of such programs; examples of poor descriptions of control treatments; variation in outcome measurements used in

individual studies; and lack of studies in the field with quantitative results suitable for meta-analysis.

Volunteers

Two reviews looked at the role of volunteers in reading development. Elbaum and Vaughn (2000) evaluated the use of one-to-one volunteers for helping 'struggling' readers and found a benefit. This review did not describe its methods for appraising the quality of included studies and it did not describe the flow of studies through the various stages to inclusion in the review. In contrast, Torgerson *et al* (2002), whilst finding a small positive effect for the use of volunteers in developing reading abilities, were more cautious as the difference between the groups was not statistically significant. Neither review assessed the potential for publication bias in the review.

Literacy instruction

One review examined the use of writing instruction in expository and narrative text (Gersten and Baker, 2001) and found a benefit. This review failed to report any raw data and inter-rater agreement on decisions throughout the review to include studies, code and quality appraise studies.

One review evaluated the effectiveness of meta-cognitive instruction in reading comprehension and found a substantial benefit (Haller *et al*, 1988). A number of meta-cognitive skills were assessed for effectiveness in three 'clusters' of mental activity (p. 6): awareness; monitoring; and regulating. Awareness of textual inconsistency was found to be the most useful strategy to be taught, and the use of self-questioning for monitoring and regulating was found to be the most beneficial strategy. This review included 20 experimental studies. However, the individual studies are not coded

according to assignment (random or otherwise) of individuals to intervention or control, and separate effect sizes are not given for RCTs and CTs. Also, as can be seen from Table 3, the reporting of methods and results was lacking in some sections of the QUORUM statement, notably in the reporting of raw data and inter-rater agreement.

Peer tutoring

A single review of peer tutoring found a positive effect (Mathes and Fuchs, 1994) compared with typical reading instruction regardless of setting. Peer tutoring was not, however, more effective when contrasted to other, teacher-led interventions, such as one-to-one teacher tutoring or teacher-led small group instruction. This review did not present raw data and did not acknowledge potential biases or draw a flow diagram of studies included and excluded at various stages of the review.

Conclusions

Although, as noted previously, the reviews tended to be on the whole of good quality, an important finding of this tertiary review is that there are relatively few systematic reviews of randomized trials in the area of literacy learning. Consequently much policy in this field cannot be underpinned with good quality evidence.

In summary, this item has examined the quality of systematic reviews in literacy learning. It has also examined how each review has dealt with possible threats to the robustness of its findings.

There are two threats to the validity of any systematic review: publication bias and poor trial quality. As noted above, a significant proportion of the reviews in this item did not

consider the issue of publication bias. The next piece of methodological work that follows on from this item is one that looks in more detail at these two threats. Therefore, in the first section of the next item the issue of publication bias is examined in detail. Following this, the threat of poor quality trials and how trial quality may affect estimates of effectiveness are examined.

References

Included systematic reviews

Bangert-Drowns, R.L. (1993) The word processor as an instructional tool: A meta-analysis of word processing in writing instruction, *Review of Educational Research*, 63(1): 69-93.

Blok, H. (1999) Reading to young children in educational settings: A meta-analysis of recent research, *Language Learning*, 49(2): 343-371.

Bus, A.G., van IJzendoorn, M.H. and Pellegrini, A.D. (1995) Joint book reading makes for success in learning to read: A meta-analysis on intergenerational transmission of literacy, *Review of Educational Research*, 65(1): 1-21.

Bus, A.G. and van IJzendoorn, M.H. (1999) Phonological awareness and early reading: A meta-analysis of experimental training studies, *Journal of Educational Psychology*, 91(3): 403-414.

Ehri, L.C., Nunes, S.R., Willows, D.M., Schuster, B.V., Yaghoub-Zadeh, Z. and Shanahan, T. (2001a) Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis, *Reading Research Quarterly*, 36(3): 250-287.

Ehri, L.C., Nunes, S.R., Stahl, S.A. and Willows, D.M. (2001b) Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis, *Review of Educational Research*, 71(3): 393-447.

Elbaum, B. and Vaughn, S. (2000) How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research, *Journal of Educational Psychology*, 92(4): 605-619.

Gersten, R. and Baker, S. (2001) Teaching expressive writing to students with learning disabilities: A meta-analysis, *Elementary School Journal*, 101(3): 251-272

Haller, E.P., Child, D.A. and Walberg, H.J. (1988) Can comprehension be taught? A quantitative synthesis of 'meta-cognitive' studies, *Educational Researcher*, 17(9): 5-8.

Jeynes, W.H. and Littell, S.W. (2000) A meta-analysis of studies examining the effect of whole language instruction on the literacy of low-SES students, *Elementary School Journal*, 101(1): 21-33.

Mathes, P.G. and Fuchs, L.S. (1994) The efficacy of peer tutoring in reading for students with mild disabilities: A best evidence synthesis, *School Psychology Review*, 23(1): 59-80.

Torgerson, C.J. and Elbourne, D. (2002) A systematic review and meta-analysis of the effectiveness of information and communication technology (ICT) on the teaching of spelling, *Journal of Research in Reading*, 25(2): 129-143.

Torgerson, C.J., King, S.E. and Sowden, A.J. (2002) Do volunteers in schools help children learn to read? A systematic review of randomized controlled trials, *Educational Studies*, 28(4): 433-444.

Torgerson, C.J. and Zhu, D. (2004) A systematic review and meta-analysis of the effectiveness of ICT on literacy learning in English, 5-16, in R. Andrews (ed.) *The Impact of ICT on Literacy Education*, London: RoutledgeFalmer.

Other references

Altman, D.G. (1996) Better reporting of randomised controlled trials: The CONSORT statement, *British Medical Journal*, 313: 570-571.

Altman, D.G., Schulz, K.F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., Gotzsche, P.C. and Lang, T. (2001) The revised CONSORT statement for

reporting randomized trials: Explanation and elaboration, *Annals of Internal Medicine*, 134(8): 663-694.

Chalmers, I., Hedges, L.V. and Cooper, H. (2002), A brief history of research synthesis, *Evaluation and the Health Professions*, 25(1): 12-37.

Cook, T.D. and Campbell, D.T. (1979) *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Boston: Houghton Mifflin Company.

Guthrie, J.T., Seifert, M. and Mosberg, L. (1983) Research syntheses in reading: Topics, audiences and citation rates, *Reading Research Quarterly*, 19(1): 16-27.

Lipsey, M.W. and Wilson, D.B. (1993) The efficacy of psychological, educational and behavioral treatment: Confirmation from meta-analysis, *American Psychologist*, 48(12): 1181-1209.

Moher, D., Cook, D.J. Eastwood, S. Olkin, I., Rennie, D. and Stroup, D.F. (1999) Improving the quality of reports of meta-analyses of randomized controlled trials: The QUORUM statement. Quality of reporting of meta-analyses, *Lancet*, 354: 1896-1900.

Shadish, W.R., Cook, T.D. and Campbell, D.T. (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Boston: Houghton Mifflin Company.

Shea, B., Dube, C. and Moher, D. (2001), Assessing the quality of reports of systematic reviews: The QUORUM statement compared to other tools, in M. Egger, G. Davey-Smith and D. Altman (eds), *Systematic Reviews in Healthcare: Meta-analysis in Context* (second edition), London: BMJ Publishing Group.

Torgerson, C.J. (2003) *Systematic Reviews*, London: Continuum Books.

Appendices

Appendix A: Summary of inclusion and exclusion criteria

Appendix B: Origin by source of all studies included in tertiary review

Appendix C: Number of studies included in tertiary review; number of studies excluded by each exclusion code

Appendix D: List of excluded studies

Appendix A: Summary of inclusion and exclusion criteria

Inclusion criteria

- Must focus on an aspect or aspects of literacy, as defined in the text
- Must be a systematic review containing a meta-analysis, as defined in the text
- Must be a systematic review and meta-analysis of experimental studies evaluating literacy interventions (with all studies containing at least one literacy outcome measure) and containing at least one randomized controlled trial, but not containing any non-experimental research
- Must focus on literacy in English as a first language (not ESL or EAL or bilingualism) and children or young people up to the age of eighteen. Where effect sizes are reported separately for different characteristics and ages the reviews are included
- Must be a systematic review, not a tertiary review

Exclusion criteria

- Not literacy (Exclude 1)
- Not a systematic review containing a meta-analysis (Exclude 2)
- Not a systematic review of literacy interventions, containing at least one RCT and no non-experimental research (Exclude 3)
- Not English as a first or sole language or not focusing on children or young people in school settings up to the age of eighteen (Exclude 4). Reviews that report effect sizes for children and adults separately are included.
- Tertiary review (Exclude 5)

Appendix B: Origin by source of all studies included in tertiary review

Source	Number of initial 'hits'	Included at first stage screening (titles and abstracts)	Unobtainable/ not received by Dec 31st 2005	Included at second stage screening (full papers)	Tertiary reviews	Systematic reviews included
PsycINFO	150	73	9	12	1	**11**
ERIC	53	16	1	0	0	**0**
C2-SPECTR	3	2	0	0	0	**0**
Contact	3	3	0	3	1	**2**
Citation	1	1	0	1	0	**1**
Total	210	95	10	16	2	**14**

Appendix C: Number of studies included in tertiary review; number of studies excluded by each exclusion code

	PsycINFO	ERIC	C2-SPECTR	Contact	Citation
Total	150	53	3	3	1
Ex 1	74	20	0	0	0
Ex 2	22	25	1	0	0
Ex 3	21	3	2	0	0
Ex 4	12	4	0	0	0
Unobtainable/not recd.	9	1	0	0	0
Ex 5	1	0	0	1	1
Systematic reviews included in review	**11**	**0**	**0**	**2**	**1**

Appendix D: List of excluded studies

Aamodt, M.G. and McShane, T.D. (1992) A meta-analytic investigation of the effect of various test item characteristics on test scores and test completion times. *Public Personnel Management*, 21(2): 151-160.

*Adams, G. and Carnine, D. Direct instruction.

Adams, C.E. (2002) Schizophrenia trials: Past, present and future. *Epidemiologia e Psichiatria Sociale*, 11(3): 144-151.

*Alexandria, V.A. (1999) Test changes: An empirical basis for defining accommodations, special education programs, Washington DC.

*Allen, P.A and Bashore, T.R. Age differences in word and language processing.

Allington, R.L. and Woodside-Jiron, H. (1997) *Adequacy of a Program of Research and of a Research Synthesis in shaping Educational Policy*, Office of Educational Research and Improvement, Washington DC, US.

*Andrews, D.A., Zinger, I., Hoge, R.D., Bonta, J. *et al.* Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis.

Apthorpe, H.S., Dean, C.B., Florian, J.E., Lauer, P.A., Reichart, R. and Snow-Renner, R. (2001) Standards in classroom practice: Research synthesis, Office of Educational Research and Improvement (ED). USA, Washington.

*Armstrong, K. (1991) Response to literature: a retrospective look at the research, University of British Columbia, Canada.

Asher, W. (1990) Educational psychology, research methodology, and meta-analysis, *Educational Psychologist*, 25(2): 143-158.

Atash, M.N. and Dawson G.O. (1986) Some effects of the ISCS program: A meta-analysis. *Journal of Research in Science Teaching*, 23(5): 377-385.

*Bain, H. (1989) A study of first grade effective teaching practices from the project star class size research. Tennessee, US.

Baldwin, R.S. and Vaughn, S. (1993) Response to Ridgeway, Dunston, & Qian: On methodological rigor: Has rigor mortis set in? *Reading Research Quarterly*, 28(4): 354-355.

*Baumeister, R.F. Writing a literature review.

Blanchard, J. and Stock, W. (1999) Meta-analysis of research on a multimedia elementary school curriculum using personal and video-game computers. *Perceptual and Motor Skills*, 88(1): 329-336.

*Boston, C. (2002) *Effect size and meta-analysis*, Office of Educational Research and Improvement, Maryland, US.

*Bradley. R., Danielson, L. and Hallahan, D.P. Identification of learning disabilities: Research to practice.

Braunger, J.L. and Lewis, J.P. (1999) Using the knowledge base in reading: Teachers at work. Northwest regional educational lab. Portland OR, USA.

Browder, D.M. and Xin, Y.P. (1998) A meta-analysis and review of sight word research and its implications for teaching functional reading to individuals with moderate and severe disabilities. *Journal of Special Education*, 32(3): 130-153.

*Buchanan, N.K. and Feldhusen, J.F. Conducting research and evaluation in gifted education: A handbook of methods and applications.

Burnett, G. (1995) Alternatives to ability grouping: still unanswered questions, Office of Educational Research and Improvement (ED). USA, Washington DC.

Burnette, J. (1999) Student groupings for reading instruction, Special education programs, Washington DC, USA.

Burns, M.K. and Symington, T. (2002) A meta-analysis of pre-referral intervention teams: Student and systematic outcomes, *Journal of School Psychology*, 40(5): 437-447.

Cabeza, R. and Nyberg, L. (1997) Imaging cognition: An empirical review of PET studies with normal subjects, *Journal of Cognitive Neuroscience*, 9(1): 1-26.

Carlson, M. and Miller, N. (1987) Explanation of the relation between negative mood and helping, *Psychological Bulletin*, 102(1): 91-108.

Carlson, M., Charlin, V. and Miller, N. (1988) Positive mood and helping behavior: A test of six hypotheses. *Journal of Personality and Social Psychology*, 55(2): 211-229.

Chabris, C.F., Steele, K.M., Bella, S.D., Peretz, I., Dunlop, T., Dawe, L.A., *et al.* (1999) Prelude or requiem for the "Mozart effect"? *Nature*, 400(6747): 826-828.

Chelimsky, E. and Shadish, W.R., editors. (1997) *Evaluation for the 21st century: A handbook.* Thousand Oaks, CA, US: Sage Publications, Inc.

Chiu, W.T. (1998) Synthesizing meta-cognitive interventions: What training characteristics can improve reading performance? Annual Meeting of American Educational Research Association, San

Diego, CA, USA, April 13-17.
Christensen, H., Hadzi-Pavlovic, D. and Jacomb, P. (1991) The psychometric differentiation of dementia from normal aging: A meta-analysis, *Psychological Assessment*, 3(2): 147-155.
Clewell, S.F. (1990) *Literacy: issues and practices*, International Reading Association, Delaware US.
Conn, V.S. and Armer, J.M. (1994) A public health nurse's guide to reading meta-analysis research reports. *Public Health Nursing*, 11(3): 163-167.
Cooper, H., Nye, B., Charlton, K. and Lindsay, J. (1996) The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research*, 66(3): 227-268.
Cortina, J.M. (2002) Big things have small beginnings: An assortment of "minor" methodological misunderstandings. *Journal of Management*, 28(3): 339-362.
Cossu, G., Rossini, F. and Marshall, J.C. (1993) Reading is reading is reading. *Cognition*, 48(3): 297-303.
*Cozby, P.C. Methods in behavioral research (4th ed.).
*Crano, W.D. and Burgoon, M. Mass media and drug prevention: Classic and contemporary theories and research.
Daneman, M. and Merikle, P.M. (1996) Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin and Review*, 3(4): 422-433.
Darkenwold, G.G. (1986) Effective approaches to teaching basic skills to adults: a research synthesis, Office of Educational Research and Improvement, Oregon US.
Davenport, L.R. (1993) The effects of homogenous groupings in mathematics, Office of Educational Research and Improvement (ED), USA, Washington DC.
Davis-Kean, P.E. (1997) A meta-analysis of preschool self-concept measures: A framework for future measures. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, 57(11-A): 4646.
*Drotar, D. Handbook of research in pediatric and clinical child psychology: Practical strategies and methods.
*Dubinsky, E., Mathews, D. and Reynolds, B.E. Readings in cooperative learning for undergraduate mathematics.
*Duer, J.M. (1998) Understanding writing disabilities: Connecticut, US.
Ehri, L.C., Shanahan, T. and Nunes, S. (2002) Response to Krashen. *Reading Research Quarterly*, 37(2):128-129.
*Elbaum, B., Vaughn, S., Hughes, M.T., Moody, S.W. and Schumm, J.S. How reading outcomes of students with disabilities are related to instructional grouping formats: A meta-analytic review.
Elbaum, B., Vaughn, S., Hughes, M. and Moody, S.W. (1999) Grouping practices and reading outcomes for students with disabilities, *Exceptional Children*, 65(3): 399-415.
Etter, J. F. (2003) Self-help smoking cessation in pregnancy, *BMJ: British Medical Journal*, 326(7386): 446-447.
Fan, W. (1994) Meta-cognition and comprehension: A quantitative synthesis of meta-cognitive strategy instruction. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, 54(10-A):3707.
Fisher, E.P. (1992) The impact of play on development: A meta-analysis. *Play and Culture*, 5(2):159-181.
Flood, J. Heath, S.B. and Lapp, D. (Eds.) (1997) *Handbook of Research on Teaching Literacy through the Communicative and Visual Arts*, Simon and Schuster, NJ, US.
Freedman, S.W. and Hechinger, F. (1992) Writing matters: Office of Educational Research and Improvement, Washington DC, US.
Froese, A.D., Gantz, B.S. and Henry, A.L. (1998) Teaching students to write literature reviews: A meta-analytic model. *Teaching of Psychology*, 25(2): 102-105.
*Fuchs, D., Fuchs, L.S., Mathes, P.G., Lipsey, M.W. and Roberts, P.H. Is "learning disabilities" just a fancy term for low achievement? A meta-analysis of reading differences between low achievers with and without the label.
*Fuchs, D., Fuchs, L.S., Mathes, P.G. and Lipsey, M.W. Reading differences between low-achieving students with and without learning disabilities: A meta-analysis.
Fukkink, R.G. and de-Glopper, K. (1998) Effects of instruction in deriving word meaning from context: A meta-analysis. *Review of Educational Research*, 68(4): 450-469.
Fusaro, J.A. (1992) Meta-analysis of the effect of sentence-combining on reading comprehension when the criterion measure is the test of reading comprehension. *Perceptual and Motor Skills*, 74(1): 331-333.
Gaffan, E.A., Tsaousis, J. and Kemp-Wheeler, S.M. (1995) Researcher allegiance and meta-analysis: The case of cognitive therapy for depression. *Journal of Consulting and Clinical Psychology*, 63(6): 966-980.
*Gersten, R.M., Schiller, E.P. and Vaughn, S. Contemporary special education research: Syntheses of the knowledge base on critical instructional issues.
Gersten, R.M. and Carnine, D. (1984) Auditory-perceptual skills and reading: A response to Kavale's

meta-analysis. *RASE: Remedial and Special Education*, 5(1): 16-19.
*Giles, D.C. Advanced research methods in psychology.
Gilger, J.W. and Pennington, B.F. (1995) Why associations among traits do not necessarily indicate their common etiology: A comment on the Geschwind-Behan-Galaburda model. *Brain and Cognition*, 27(1):89-93.
Goetz, E.T. and Sadoski, M. (1995) "Commentary: The perils of seduction: Distracting details or incomprehensible abstractions?": Reply. *Reading Research Quarterly*, 30(3): 518-519.
Goldring, E.B. and Presbrey, L.S. (1986) Evaluating preschool programs: A meta-analytic approach. *Educational Evaluation and Policy Analysis*, 8(2): 179-188.
*Graham, S. and Harris, K.R. Students with learning disabilities and the process of writing: A meta-analysis of SRSD studies.
*Grimm, L.G. and Yarnold, P.R. Reading and understanding multivariate statistics.
Gunn, B. K., *et al.* (1995) Emergent literacy: curricular and instructional implications for diverse learners, National Centre to Improve the Tools for Educators, Oregon, US.
Guthrie, J.T., Seifert, M. and Mosberg, L. (1983) Research syntheses in reading: Topics, audiences and citation rates, *Reading Research Quarterly*, 19(1): 16-27.
Guzzetti, B.J. (2000) Learning counter-intuitive science concepts: What have we learned from over a decade of research? *Reading and Writing Quarterly: Overcoming Learning Difficulties*, 16(2): 89-98.
Guzzetti, B.J., Young BJ. Peyton J, Gritsavage, M.M., Fyfe L.M. and Hardbrook, M. (2002) Reading, writing and talking gender in literacy learning, Literacy Studies Series. USA, Delaware.
Haigh, R. (2002) Therapeutic community research: Past, present and future. *Psychiatric Bulletin*, 26(2): 65-68.
Hall, K. (2002) Developing the protocol for a systematic review of literature on effective literacy teachers and their teaching. *Reading literacy and language*, 36(1): 44-47.
Hall, T.E., Hughes, C.A. and Filbert, M. (2000) Computer assisted instruction in reading for students with learning disabilities: A research synthesis. *Education and Treatment of Children*, 23(2): 173-193.
Harris, C., Kelly, C., Valentine, J.C. and Muhlenbruck, L. (2000) Making the most of summer school: A meta-analytic and narrative review. *Monographs of the Society for Research in Child Development*, 65(1): v-118.
Harste, J.C. (1993) Response to Ridgeway, Dunston, & Qian: Standards for instructional research. *Reading Research Quarterly*, 28(4): 356-358.
Hartley, J. and McKeachie, W.J., editors. (1990) *Teaching psychology: A handbook: Readings from "Teaching of Psychology."* Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.
Haury, D.L. and Milbourne, L.A. (1999) Should students be tracked in math or science? Office of Educational Research and Improvement (ED), USA, Washington DC.
Hawisher, G.E. and Self, C.L. (1994) Bibliography of composition and rhetoric, Southern Illinois University Press. Illinois, US.
Henk, W.A. and Stahl, N.A. (1985) A meta-analysis of the effect of notetaking on learning from lecture, *National Reading Conference Yearbook*, (34): 70-75.
Heubusch, J.D. and Lloyd, J.W. (1998) Corrective feedback in oral reading. *Journal of Behavioral Education*, 8(1): 63-79.
Hiebert, E.H. (1987) The context of instruction and student learning: an examination of Slavin's assumptions. *Review of Educational Research*, 57: 337-340.
*Hiebert, R.E. Public Relations Review: A journal of research and comment. 200: 26.
Hillocks, G. (1984) What works in teaching composition: A meta-analysis of experimental treatment studies. *American Journal of Education*, 93(1): 133-170.
Hollifield, J. (1987) Ability grouping in elementary schools, Office of Educational Research and Improvement. Washington DC, US.
Horn, W.F. and Packard, T. (1985) Early identification of learning problems: A meta-analysis. *Journal of Educational Psychology*, 77(5): 597-607.
Hutchinson, L., Aitken, P. and Hayes, T. (2002) Are medical postgraduate certification processes valid? A systematic review of the published evidence, *Medical Education*, 36(1): 73-91.
Jesen, J.W. and Winitzky, N. (1999) What works in teacher education? Annual meeting of the American Association of Colleges for Teacher Education, USA Washington DC, Feb 24-27.
Josephs, R.A. and Hahn, E.D. (1995) Bias and accuracy in estimates of task duration, *Organizational Behavior and Human Decision Processes*, 61(2): 202-213.
Kamil, M. L., Mosenthal, P.B., Pearson, P.D. and Barr, R., editors. (2000) *Handbook of reading research.* Mahwah, NJ, US: Lawrence Erlbaum Associates, Publishers.
Kaplan, H., *et al.* (1993) Research synthesis on design of effective media, materials and technology for deaf and hard of hearing students, National Centre to Improve the Tools for Educators, Oregon, US.

Kardash, L.A.M. and Wright, L. (1987) Does creative drama benefit elementary students: A meta-analysis, *Youth-Theatre-Journal*, 1(3): 11-18.

Katz, L.G. (2000) Academic redshirting and young children, Office of Educational Research and Improvement (ED). USA, Washington DC.

Kavale, K.A. and Forness, S.R. (2000) Policy decisions in special education: The role of meta-analysis. In: Gersten, R.M. and Schiller, E.P., editors. *Contemporary special education research: Syntheses of the knowledge base on critical instructional issues, The LEA series on special education and disability*. Mahwah, NJ, US: Lawrence Erlbaum Associates, Publishers, 281-326.

Kavale, K.A. (1980) Auditory-visual integration and its relationship to reading achievement: A meta-analysis. *Perceptual and Motor Skills*, 51(3, Pt 1): 947-955.

Kavale, K.A. (1981) The relationship between auditory perceptual skills and reading ability: A meta-analysis. *Journal of Learning Disabilities*, 14(9): 539-546.

Kavale, K.A. (1982) A meta-analysis of the relationship between visual perceptual skills and reading achievement, *Journal of Learning Disabilities*, 15(1): 42-51.

Kavale, K.A. (1984) A meta-analytic evaluation of the Frostig test and training program, *Exceptional Child*, 31(2): 134-141.

Kavale, K.A. and Forness, S.R. (1984) A meta-analysis of the validity of Wechsler Scale profiles and recategorizations: Patterns or parodies? *Learning Disability Quarterly*, 7(2): 136-156.

Kavale, K.A. and Forness, S.R. (2000) Auditory and visual perception processes and reading ability: A quantitative reanalysis and historical reinterpretation. *Learning Disability Quarterly*, 23(4): 253-270.

*Kazdin, A.E., Methodological issues & strategies in clinical research (3rd ed.).

*Kazdin, A.E., Bem, D.J., Maxwell, S.E., Cole, D.A, Rosenthal, R. and Oetting, E.R. Publication and communication of research.

*Keeves, J.P., Educational research, methodology, and measurement: An international handbook.

Kinsey, S.J. (2002) La agrupación de edades multiples y el logro académico, Office of Educational Research and Improvement, Washington DC.

*Kirsch, I. (1998) On the importance of reading carefully: A response to Klein, *Prevention and Treatment*, 1.

Klauer, K.J. (1984) Intentional and incidental learning with instructional texts: A meta-analysis for 1970-1980. *American Educational Research Journal*, 21(2): 323-339.

Klesius, J.P. and Searls, E.F. (1990) A meta-analysis of recent research in meaning vocabulary instruction, *Journal of Research and Development in Education*, 23(4): 226-235.

Kliewer, C. and Biklen, D. (2001) "School's not really a place for reading": A research synthesis of the literate lives of students with severe disabilities, *Journal of the Association for Persons with Severe Handicaps*, 26(1): 1-12.

Klumb, K. (1990) Patterns of children's reading and spelling substitution errors, Lucerne Valley Unified School District, California, US.

Kroenke, K., Taylor-Vaisey, A., Dietrich, A.J. and Oxman, T.E. (2000) Interventions to improve provider diagnosis and treatment of mental disorders in primary care: A critical review of the literature. *Psychosomatics: Journal of Consultation Liaison Psychiatry*, 41(1): 39-52.

*Leong, F.T.L. and Austin, J.T. The psychology research handbook: A guide for graduate students and research assistants.

Liao, Y.K.C. and Bright, G.W. (1991) Effects of computer programming on cognitive outcomes: A meta-analysis. *Journal of Educational Computing Research*, 7(3): 251-268.

Lipsey, M.W. and Wilson, D.B. (1993) The efficacy of psychological, educational and behavioral treatment: Confirmation from meta-analysis, *American Psychologist*, 48(12): 1181-1209.

Ludlow, L.H. (1987) The graphical representation of quantitative research synthesis residual variation. *Educational and Psychological Measurement*, 47(4): 941-951.

Lysynchuk, L.M., Pressley, M., d'Ally, H., Smith, M. and Cake, H. (1989) A methodological analysis of experimental studies of comprehension strategy instruction, *Reading Research Quarterly*, 24(4): 458-470.

MacArthur, C.A., Ferretti, R. P., Okolo, C. M. and Cavalier, A. R. (2001) Technology applications for students with literacy problems: A critical review. *Elementary School Journal*, 101(3) 273-301.

Madamba, S. R. (1981) Meta-analysis on the effects of open and traditional schooling on the teaching and learning of reading, *Dissertation Abstracts International*, 41(8-A): 3508.

Marmolejo, A. (1990) The effects of vocabulary instruction with poor readers: A meta-analysis. *Dissertation Abstracts International*, 51(3-A): 807.

Maxwell, L. (1986) Making the most of ability groupings: research in brief. Office of Educational Research and Improvement, Washington DC, US.

McEvoy, T. J. (1983) A meta-analysis of comparative research on the effect of desegregation on academic achievement and self-esteem of Black students. *Dissertation Abstracts International*, 43(11-A): 3559.

McGiverin, J., Gilman, D. and Tillitski, C. (1989) A meta-analysis of the relation between class size and achievement. *Elementary School Journal*, 90(1): 47-56.

McManus, I. C., Porac, C., Bryden, M. P. and Boucher, R. (1999) Eye-dominance, writing hand, and throwing hand. *Laterality*, 4(2): 173-192.

Metsala, J. L., Stanovich, K. E. and Brown, G. D. A. (1998) Regularity effects and the phonological deficit model of reading disabilities: A meta-analytic review. *Journal of Educational Psychology*, 90(2): 279-293.

Midence, K., McManus, C., Fuggle, P. and Davies, S. (1996) Psychological adjustment and family functioning in a groups of British children with sickle cell disease: Preliminary empirical findings and a meta-analysis. *British Journal of Clinical Psychology*, 35(3): 439-450.

Miller, N., Lee, J. Y. and Carlson, M. (1991) The validity of inferential judgments when used in theory-testing meta-analysis. *Personality and Social Psychology Bulletin*, 17(3): 335-343.

Mullen, B., Johnson, C. and Salas, E. (1991) Productivity loss in brainstorming groups: A meta-analytic integration. *Basic and Applied Social Psychology*, 12(1): 3-23.

*Musthafa, B. (1995) *Play-Literacy Connections: a Research Synthesis and Suggested Directions*, Ohio, US

Myers, D. G. (1991) Union is strength: A consumer's view of meta-analysis. *Personality and Social Psychology Bulletin*, 17(3): 265-266.

National Center for ESL literacy education (2002) Proceedings of the national symposium on adult ESL research and practice, Washington DC, US.

National Assessment of Educational Progress (1995) A synthesis of data from NEAP's 1992 integrated reading record at grade 4. Centre for Assessment of Educational Progress, Washington DC, US.

Neilsen, L. (1993) Response to Ridgeway, Dunston, & Qian: Authoring the questions: Research as an ethical enterprise. *Reading Research Quarterly*, 28(4): 350-353.

Neville, D. D. and Searls, E. F. (1991) A meta-analytic review of the effect of sentence-combining on reading comprehension. *Reading Research and Instruction*, 31(1): 63-76.

*Nietzel, M. T., Bernstein, D. A. and Milich, R. *Introduction to clinical psychology* (3rd ed.).

Northwest Regional Educational Lab (1990) School Improvement Research Series 4. Office of Educational Research and Improvement, Oregon US.

O'Shaughnessy, T. E. and Swanson, H. L. (1998) Do immediate memory deficits in students with learning disabilities in reading reflect a developmental lag or deficit? A selective meta-analysis of the literature. *Learning Disability Quarterly*, 21(2): 123-148.

Obrzut, J. E., Boliek, C. A. and Bryden, M. P. (1997) Dichotic listening, handedness, and reading ability: A meta-analysis. *Developmental Neuropsychology*, 13(1): 97-110.

Office of Educational Research and Improvement (1993) *Advances in Educational Research* Vol 1. Washington, DC, US.

Patten, P. and Ricks OB. (2000) *Childcare quality: and overview for parents*. Office of Educational Research and Improvement (ED). USA, Washington DC.

*Phillips, M., Crouse, J. and Ralph, J. Does the Black-White test score gap widen after children enter school?

*Pierce, P.L. (1994) *Technology Integration into Early Childhood Curricula: Where We've Been, Where We Are, Where We should go*, North Carolina, US.

Powers, S. and Rossman, M. H. (1984) Evidence of the impact of bilingual education: A meta-analysis. *Journal of Instructional Psychology*, 11(2): 75-78.

*Ramsay, S. PR Bibliography, Public Relations Review, 24, 1998.

*Ramsay, S. PR Bibliography, Public Relations Review, 25, 1999.

Reading, M. and McDowell, F. H. (1994) Stroke rehabilitation outcome studies: Selection for meta-analysis. *Archives of Neurology*, 51(2): 120.

Reading, R. (2002) Safety education of pedestrians for injury prevention: A systematic review of randomized controlled trials. *Child Care, Health and Development*, 28(5): 432.

Reay, D. G., Harrison, G. V. and Gottfredson, C. (1984) The effect on pupil reading achievement of teacher compliance with prescribed methodology. *Research in Education*, (32): 17-23.

*Reinwein, J. and Huberdean, L. (1997) A second look at Dwyer's studies by means of meta-analysis: The effects of pictorial realism in text comprehension and vocabulary.

*Reise, S. P. and Duan, N. Multilevel modeling: Methodological advances, issues, and applications.

*Renninger, K. A., Hidi, S. and Krapp, A. The role of interest in learning and development.

Ridgeway, V. G., Dunston, P. J. and Qian, G. (1993) A methodological analysis of teaching and learning strategy research at the secondary school level. *Reading Research Quarterly*, 28(4): 334-349.

*Rogelberg, S. G. Handbook of research methods in industrial and organizational psychology.

Rose, S., Bisson, J. and Wessely, S. (2003) A systematic review of single-session psychological interventions ('debriefing') following trauma. *Psychotherapy and Psychosomatics*, 72(4): 176-184.

Rosenthal R. (1995) Writing meta-analytic reviews. *Psychological Bulletin*, 118(2): 183-192.

*Rosnow, R. L. and Rosenthal, R. Beginning behavioral research: A conceptual primer (3rd ed.).
Ross, S. (1998) Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing*, 15(1): 1-20.
Russell V.J., Rowe, K.J. and Hill, P.W. (1998) Effects of multigrade classes on student progress in literacy and numeracy: Quantitative evidence and perceptions of teachers and school leaders. Annual meeting of the Australian Association for Research in Education, Adelaide, Australia.
*Salkind, N. J. Exploring research.
Schafer, W.D. (2001) Replication in field research. Office of Educational Research and Improvement. USA, Washington DC.
Schutte, J.W. and Hosch, H. M. (1997) Gender differences in sexual assault verdicts: A meta-analysis. *Journal of Social Behavior and Personality*, 12(3): 759-772.
*Schwarzer, R., Van-der-Ploeg, H. M. and Spielberger, C. D. Advances in test anxiety research, Vol. 5.
Scope, E.E. (1999) A meta-analysis of research on creativity: The effects of instructional variables. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, 59(7-A): 2348.
*Shadish, W.R. and Fuller, S. The social psychology of science.
Short, K.G. (1995) Professional Resources in Children's literature: Piecing a patchwork quilt. International Reading Association, Delaware, US.
Simmons, D.C. and Kameenui, E. J., editors. (1998) *What reading research tells us about children with diverse learning needs: Bases and basics*. Mahwah, NJ, US: Lawrence Erlbaum Associates, Publishers.
Slavin, R.E., *et al.* (1994) Preventing early school failure: research policy and practice. California alliance for elementary education. Mass, US.
Slavin, R.E. (1986) Ability grouping and student achievement in elementary schools: a best evidence synthesis. Centre for Research on Elementary and Middle Schools, Maryland, US.
*Smith, M. and Robertson, I. T. Advances in selection and assessment.
*Smith, S.B., *et al.* Synthesis of research on phonological awareness: principles and implications for reading acquisition. National Centre to improve the tools for educators, Oregon, US, 1995.
Smyth, J.M. (1998) Written emotional expression: Effect sizes, outcome types, and moderating variables. *Journal of Consulting and Clinical Psychology*, 66(1): 174-184.
Stahl, S.A. and Miller, P.D. (1989). Whole language and language experience approaches for beginning reading: a quantitative research synthesis. *Review of Educational Research*, 59(1): 87-116.
Stahl, S.A. and Fairbanks, M.M. (1986) The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research*, 56(1): 72-110.
Stahl, S.A., McKenna, M.C. and Pagnucco, J.R. (1994) The effects of whole-language instruction: An update and a reappraisal. *Educational Psychologist*, 29(4): 175-185.
Stone, C.L. (1983) A meta-analysis of advance organizer studies. *Journal of Experimental Education*, 51(4): 194-199.
Stuebing, K.K., Fletcher, J.M., LeDoux, J.M., Lyon, G.R., Shaywitz, S.E. and Shaywitz, B.A. (2002) Validity of IQ-discrepancy classifications of reading disabilities: A meta-analysis. *American Educational Research Journal*, 39(2): 469-518.
Swanborn, M.S.L. and Glopper, K.D. (1999) Incidental word learning while reading: A meta-analysis. *Review of Educational Research*, 69(3): 261-285.
*Swanson, H. L., Harris K-R. and Graham, S. *Handbook of learning disabilities*.
Swanson, H.L. (1999) Reading research for students with LD: A meta-analysis in intervention outcomes. *Journal of Learning Disabilities*, 32(6): 504-532.
Swanson, H.L. and Sachse-Lee, C. (2000) A meta-analysis of single-subject-design intervention research for students with LD. *Journal of Learning Disabilities*, 33(2): 114-136.
Swanson, H.L. (2001) Research on interventions for adolescents with learning disabilities: A meta-analysis of outcomes related to higher-order processing. *Elementary School Journal*, 101(3): 331-348.
Thompson, C. (2001) "Declarations of interest': Reply. *British Journal of Psychiatry*, 179:175.
*Tracy, D.H. (1995) Family literacy: research synthesis. New Jersey US.
Troia, G.A. (1999) Phonological awareness intervention research: A critical review of the experimental methodology, *Reading Research Quarterly*, 34: 28-52.
Tyler, J.H. (2002) The economic benefits of the GED: a research synthesis. NCSALL Research Brief. National Centre for the Study of learning and literacy, Boston Ma, USA.
Van Broekhuizen, D.L. (2000) Literacy in indigenous communities. Research Series. Office of Educational Research and Improvement (ED), USA, Washington DC.
Van IJzendoorn, M.H. and Bus, A.G. (1994) Meta-analytic confirmation of the nonword reading deficit in developmental dyslexia. *Reading Research Quarterly*, 29(3): 266-275.
*Walker, A. (Ed.) (1993) PR Bibliography, Public Relations Review, 19.
*Walker, A. (Ed.) (1994) PR Bibliography, Public Relations Review, 20.
*Walker, A. (Ed.) (1995) PR Bibliography, Public Relations Review, 21.

*Walker, A. (Ed.) (1997) PR Bibliography, Public Relations Review, 23.
Wagner, R.K. (1988) Causal relations between the development of phonological processing abilities and the acquisition of reading skills: A meta-analysis. *Merrill Palmer Quarterly*, 34(3): 261-279.
Welch, C.J. and Miller, T.R. (1995) Assessing differential item functioning in direct writing assessments: Problems and an example. *Journal of Educational Measurement*, 32(2): 163-178.
*Wentland, E.J. Survey responses: An evaluation of their validity.
Wilcock, G.K., Birks, J., Whitehead, A. and Evans, S.J.G. (2002) The effect of selegiline in the treatment of people with Alzheimer's disease: A meta-analysis of published trials. *International Journal of Geriatric Psychiatry*, 17(2): 175-183.
Willig, A.C. (1985) A meta-analysis of selected studies on the effectiveness of bilingual education. *Review of Educational Research*, 55(3): 269-317.
*Wilsher, C.R. Pharmacological treatment of dyslexia.
Wong, B.Y.L. (2001) Commentary: Pointers for literacy instruction from educational technology and research on writing instruction. *Elementary School Journal*, 101(3): 359-369.
Woodward, M., Nursten, J., Williams, P. and Badger, D. (2000) Mental disorder and homicide: A review of epidemiological research. *Epidemiologia e Psichiatria Sociale*, 9(3): 171-189.
*Wasik, B.A. and Slavin, R.E. (1990) Preventing early reading failure with one-to-one tutoring: A best-evidence synthesis.

*References marked with an asterisk are incomplete. If a study had an incomplete citation but could confidently be excluded using the title and abstract only, the full citation was not sought. However, if any study with an incomplete citation could not confidently be excluded using the information given in the imported citation, the full citation was sought, and if necessary the paper was obtained to check for inclusion/exclusion. In Appendix D the full citation as obtained from the electronic searches is given for each marked study.

Item 3:

Methodological issues based on systematic reviews

Contents

Section 1:

Introduction

Chapter 1: Introduction to the methodological issues: publication bias and design bias

In the preceding item (Item 2), a tertiary review was undertaken of systematic reviews and meta-analyses of randomized controlled trials (RCTs) in literacy learning. Although systematic reviews are much less likely to produce biased results than narrative reviews, they are only as valid as the sample of trials they identify and the quality of the trials they include. Two important issues can affect the validity of systematic reviews: publication bias and design bias.

The structure of this following item is as follows. Section 2 examines the problem of trial identification for inclusion in systematic reviews: publication bias. Section 3 examines the problem of design quality of individual trials included in systematic reviews: design bias. The rationale for this approach is: publication and design bias are the two main sources of bias in any systematic review. If a biased sample of trials is identified, the results and conclusions can mislead (publication bias). Even if all the trials in a field are identified, bias can still result if some or all of these trials have design weaknesses (design bias).

The purpose of this item is to assess how publication bias and design bias may have affected systematic reviews in literacy learning. The item contains substantive methodological work using, mainly, some of the data generated from the tertiary review in the previous item. This item makes two important and distinct contributions to the literature. Firstly, a range of methods is applied to the sample of systematic reviews in literacy learning to see if it is susceptible to publication bias. This includes a case study replication, update and secondary analysis of one review, which involves a search for

any relevant unpublished literature that may have been excluded from the original review and quality appraisal of all the included RCTs. Secondly, the methodological reliability of the individual randomized trials in the systematic reviews is assessed, and an exploratory meta-regression is undertaken to look at whether there is any relationship between study quality (such as sample size and blinding of outcome assessment) and effect size. As far as I am aware, the systematic application of these methodological techniques to a range of meta-analyses in literacy research has not previously been undertaken.

Section 2: Publication bias

The results of a systematic review are valid if either all the relevant trials are identified and included, or if a representative sample of studies is included. Selection of an unrepresentative sample can lead to misleading results. Many trials are undertaken and, for one reason or another, not published. It is possible that the results of the unpublished trials may differ from the results of the published studies. If this is the case then any systematic review based upon the published trials could be biased.

In Chapter 2 the history of publication bias is explored in order to locate the present study within the context of previous methodological work. The issue of publication bias was recognised by social science researchers more than 40 years ago. The chapter discusses a number of ways in which publication bias can be identified within an individual review. In addition, the chapter describes methods of adjusting the findings of a systematic review in the light of probable publication bias.

Chapter 3 assesses the extent to which the 14 systematic reviews in literacy learning identified in the tertiary review (Item 2 of the portfolio) included fugitive studies. It investigates the extent of possible publication bias in the field of literacy learning by comparing the effect sizes and sample sizes between published and unpublished literature. A sub-sample of systematic reviews identified in the tertiary review is examined in depth using funnel and normal quantile plots to assess whether or not there is evidence for publication bias. It also compares the findings of each of these methods in order to check for reliability. The chapter concludes by highlighting reviews where there is evidence of publication bias and recommends caution in interpreting their results.

Chapter 4 is a case study of one of the systematic reviews where possible publication bias was identified. The original searches are replicated in the grey literature and one unpublished trial not included in the original review is included. The review is updated, and relevant trials published since 2000 are included. The chapter describes an analysis comparing the updated review with the original. A new funnel plot and a normal quantile plot are presented, which show where the new trials are located. Issues of internal validity and generalisability uncovered in the process of replicating and updating the review are highlighted. These are discussed in the light of the quality appraisal undertaken in the tertiary review, using an adaptation of the QUORUM statement.

Section 3: Design bias

The quality of included trials can affect the results of systematic reviews. Methodological work in healthcare research has noted that poor quality trials may

exaggerate the effectiveness of an intervention (Schulz *et al*, 1995). However, there has been little similar methodological work in the field of systematic reviews in education.

Chapter 5 discusses a number of the important issues with respect to the quality of RCTs. It also discusses how educational trialists can avoid some of the potential biases that might occur when conducting a trial.

Chapter 6 provides an overview of the quality issues with respect to the design of trials. A large number of quality appraisal tools have been developed, mainly in the context of healthcare trials. In Chapter 6, a healthcare quality checklist is adapted for use in educational trials and is compared with a checklist developed by an educational researcher. These tools are compared using a methodological review identified through the process of undertaking the tertiary review (Troia, 1999). This chapter then goes on to examine the prevalence of some of the key quality items in a wider range of educational trials and looks to see if the quality of recent trials is better than the quality of older trials.

After Chapter 6 has set the scene of design issues, Chapter 7 focuses on individual design issues. One source of bias that is examined in Chapter 7 is 'contamination' or 'leakage' between treatment and control groups. Because educational interventions often have a higher possibility to leak out to the control group than many healthcare trials, this may result in dilution of a treatment effect. One remedy for this is to randomize pupils or students in groups (e.g., by school or class). In Chapter 7 the effect sizes of group or cluster trials are compared with the effect sizes of individual trials. The hypothesis tested is that group allocated trials will tend to have larger effect sizes than individually randomized studies. The chapter also examines the unique

methodological issues with respect to cluster-randomized studies, for example, the use of appropriate statistical techniques

Blinded follow-up of trial participants has been noted as a crucial methodological issue for many years (Cook and Campbell, 1979; Shadish *et al*, 2002). In Chapter 7 the issue of blinded follow-up of trial participants is explored. Trials are grouped into those that used blinded follow-up and those that did not. A meta-analysis comparing the two groups of trials is undertaken to see if this has an effect on the size of the trial outcome.

Quality appraisal of trials is difficult, not least because many trialists do not report their methods in sufficient detail to make an assessment. One proxy measure of quality is that of sample size. Kjaergard (2001), in the context of healthcare trials, found that small trials tended to be methodologically poor. Assuming this phenomenon applies to educational trials, Chapter 7 correlates the effect sizes of the trials with their sample sizes. The hypothesis is that smaller trials tend to have larger effect sizes. In addition, the quality of small trials is directly compared with the quality of larger trials in areas such as blinding, attrition and intention to teach analysis. Finally, an exploratory meta-regression analysis is undertaken, to assess in a multivariate fashion which quality items are the most important predictors of effect size. All of the trials that included effect sizes and quality indicators have been entered into the analysis.

Section 2:

Publication bias

Chapter 2: Publication bias in systematic reviews of randomized controlled trials[7]

Introduction

Systematic reviewing presents a transparent and replicable approach to locating, identifying and synthesizing all the research literature in any given field. Systematic reviews aim either to exhaustively search for a population of studies or to sample representatively from it (Smith, 1980; Torgerson, 2003). Two main potential threats to the validity of systematic reviews are reviewer selection bias and publication bias. Systematic review methods reduce the problem of reviewer selection effects. Reviewer selection effects occur when the criteria for study inclusion are developed in such a way as to 'select' into the review a biased sample of published studies. Because systematic reviews use transparent and replicable inclusion and exclusion criteria they are less likely to be affected by biased or selective reporting of the research literature than traditional narrative reviews. Despite this, however, the results of systematic reviews can be biased if there is a significant problem with publication bias. Researchers have long suggested that the published studies in the social sciences represent a biased sample of all the studies that are carried out (Rosenthal, 1979; Smith, 1980). Rosenthal described an extreme view of the problem:

> '..the journals are filled with the 5% of the studies that show type I errors, while the file drawers back at the lab are filled with the 95% of the studies that show non-significant (e.g. $p>0.05$) results' (Rosenthal, 1970, p.638).

Publication bias is one of a range of reporting biases (including language bias and citation bias) that can affect the results of systematic reviews and meta-analyses of trials (true or quasi-experiments) and has been widely reported in the methodological literature (Sterne *et al*, 2000). Also known as positive outcome bias or the file-drawer

[7] This chapter will be published as: Torgerson, C.J. (2006) 'Publication bias: The Achilles' heel of systematic reviews?' in *British Journal of Educational Studies*, 54(1). See Appendix A.

effect, publication bias refers to the tendency for a greater proportion of statistically significant positive results of experiments to be published and, conversely, a greater proportion of statistically significant negative or null results not to be published (Greenwald, 1975; Rosenthal, 1979; Hedges and Olkin, 1980; Light, 1983; Light and Pillemer, 1984; Dickersin *et al*, 1987; Iyengar and Greenhouse, 1988; Begg and Berlin, 1988; Dickersin, 1997; Dickersin, 2002; Fitz-Gibbon, 2004). It also manifests as the tendency for published studies to have higher effect sizes than unpublished studies (Smith, 1980; Kulik and Kulik, 1989; Durlak and Lipsey, 1991), and for published studies to have larger sample sizes. The smaller the study, the larger will be the intervention effect necessary to demonstrate a statistically significant effect (Lipsey and Wilson, 2001). Therefore publication bias, if present in a review, will be partly a function of sample size (Dear and Begg, 1997), and a meta-analysis containing a large number of small studies will have an increased risk of publication bias (Begg and Berlin, 1988). Publication bias can be due either to researchers tending not to submit their non-significant results and/or to journal editors tending not to accept them for publication (Lipsey *et al*, 1985; Wilson and Lipsey, 2001; Begg and Berlin, 1988).

If publication bias exists in a field, researchers searching for potentially relevant studies to include in systematic reviews will find studies with significant positive results easier to retrieve than studies with significant negative results (Wilson and Lipsey, 2001). If positive results are more likely to be published, this will bias the review towards a positive result because published studies are likely to be 'over-represented' in systematic reviews (Iyengar and Greenhouse, 1988; Lipsey and Wilson, 1993; Wilson and Lipsey, 2001; Smith, 1980). Publication bias is, therefore, a potentially major threat to the validity of systematic reviews. Note, however, that publication bias also

affects non-systematic reviews in addition to their, usually unacknowledged, identification bias.

History of publication bias

Many researchers have demonstrated that the problem of publication bias is widespread. There is a consensus that the problem exists and that it is serious (Begg and Berlin, 1988). Selective publication was identified as being a problem in meta-analyses of experimental studies in educational research over 40 years ago (Sterling, 1959; Smart, 1964).

I have undertaken an overview of the effects of publication bias in the literature. There are many references to publication bias in the healthcare literature. Although I have included some key articles from healthcare research, I have deliberately kept my overview mainly in the non-healthcare literature, where possible.

The three earliest studies that looked at publication bias I identified were in the psychological literature. Smart, Cohen and Sterling demonstrated that the majority of published studies in the field of psychology had statistically significant findings (Sterling, 1959; Cohen, 1962; Smart, 1964).

In his early cross-sectional study Sterling (1959) demonstrated that, in four major psychological journals published in 1955 and 1956, there was a greater probability of the results of experiments being published if the relevant test of significance rejected the null hypothesis than if the test failed to reject the hypothesis. In order to demonstrate that research that yielded non-significant results was routinely not being published,

Sterling searched all the issues of four journals for the period January to December of either 1955 or 1956. Out of a total of 294 articles, 286 rejected the null hypothesis (at the 0.05 level of significance) and only eight failed to reject the null hypothesis at this level of significance. Sterling concluded that, because research yielding non-significant results was not being published, such research could be repeated until eventually, by chance, a significant result would occur (a Type 1 error) and would consequently be published, leading to erroneous conclusions about the effectiveness of the intervention.

A few years after Sterling, Cohen (1962) surveyed the *Journal of Abnormal and Social Psychology* for the two years 1960 and 1961. He analysed the 70 articles that involved major statistical tests for their power to detect small, medium and large effects using 2-tailed tests (at the 0.05 level of significance). He found that the mean power values (i.e., the probability of rejecting false null hypotheses), over the 70 empirical studies, were: 0.18 for small effects; 0.48 for medium effects; and 0.83 for large effects. Because virtually all of the trials in Cohen's review had found, as statistically significant, small to medium effect sizes, unpublished studies with non-significant findings must have been missing. Cohen concluded that the power of these studies was far too small (unless the effect sought was large) and had probably led to the failure to reject false null hypotheses. As published research under-represents the research undertaken in a field, it is likely that significant numbers of studies having non-statistically significant findings had not been published in this journal for the two years surveyed.

In another cross-sectional study which replicated Sterling's work, Smart (1964) demonstrated that unpublished studies (conference papers) contained a higher proportion of negative results than did studies published in psychological journals; and

that PhDs in psychology that reported negative results were less likely to be published than were those with positive results. He coded and compared: all of the psychological experiments published in four journals in 1961 and 1962; a non-random sample of 100 PhD theses from 1962; and a random sample of 100 unpublished papers presented at the *American Psychological Association* in 1962. He concluded that the neglect of negative studies was due to non-submission by authors or to greater critical examination of experiments containing negative results by journal editors and peer reviewers.

In order to estimate bias against the null hypothesis, Greenwald (1975) surveyed the authors and reviewers of all the manuscripts processed by him as associate editor of the *Journal of Personality and Social Psychology* during a three-month period in 1973. He asked these authors and reviewers about the relative probability of submitting studies for publication that either rejected or accepted the null hypothesis. The results indicated a strong bias against accepting the null hypothesis, illustrated by the 0.49 probability of submitting a rejection of the null hypothesis for publication compared with the low probability of 0.06 for submitting an acceptance of the null hypothesis for publication. Greenwald confirmed these findings by examining every article published in the *Journal of Personality and Social Psychology* in the year 1972 to determine what proportion accepted the null hypothesis. He found that out of a total of 100 articles only 24 reported acceptance of the null hypothesis.

In 1980, Smith examined a sub-sample of 12 meta-analyses in the fields of educational, social and psychological research and found that the findings from journals were, on average, one-third of a standard deviation more skewed towards the rejection of the null hypothesis than findings reported in theses or dissertations: that is a mean effect size of

0.64 in the published literature compared with a mean effect size of 0.48 in the unpublished literature (Smith, 1980).

These early findings have been more recently confirmed in the fields of healthcare research (Dickersin, 2002) and psychological, educational and behavioural treatment research (Lipsey and Wilson, 1993).

In their tertiary review of meta-analyses of psychological, educational and behavioural treatment research, Lipsey and Wilson (1993) found a 'strong skew' towards positive effects. They included 302 meta-analyses in their review. Only 6 of these reported negative effect sizes and relatively few reported effect sizes around zero – 85% of all the effect sizes were greater than 0.2. Lipsey and Wilson tried to identify reasons for this positive skew in their data and concluded that there were a number of possible factors leading to bias: selection bias and publication bias. They looked to see if evidence for the latter could explain the strongly positive effects by examining the differences in effect sizes between published and unpublished studies. They analysed a subset of 92 meta-analyses that reported separate effect sizes for published and unpublished studies, and found that the published studies had mean effect sizes 0.14 standard deviations larger than the mean effect sizes of the unpublished studies. These data support the view that studies with larger effect sizes are more likely to be published because, other things being equal, they will be more likely to be statistically significant.

In 1995, Sterling *et al* replicated Sterling's earlier study of the percentage of published articles in four major psychology journals that rejected the null hypothesis. This time Sterling and colleagues (1995) looked at eight psychology journals and three medical journals for either 1986 or 1987. They found that publication patterns in 1986/7 were

still consistent with publication bias and that there had been little change since the original study. In the eight psychology journals, 95.56% of articles using tests of significance rejected the null hypothesis (compared with 97.28% in 1958). The authors concluded that the practice of psychology journals preferring positive to negative results had not changed over the thirty-year period between 1956 and 1986. In the paper the authors cite a letter from an editor of a major environmental/toxicological journal explaining why a manuscript had been rejected:

> Unfortunately, we are not able to publish this manuscript. The manuscript is very well written and the study was well documented. Unfortunately, the negative results translate into a minimal contribution to the field. We encourage you to continue your work in this area and we will be glad to consider additional manuscripts that you may prepare in the future (cited in Sterling *et al*, 1995, p.109).

Kulik and Kulik examined four of their own meta-analyses in the field of educational research for evidence of publication bias (Kulik and Kulik, 1989). The meta-analyses were undertaken in the areas of computer-based instruction at elementary and secondary level; computer-based instruction at post-secondary level; ability grouping; and mastery learning systems. Kulik and Kulik compared the mean effect sizes in all four meta-analyses for unpublished reports, unpublished dissertations and published journal articles. They found that in two out of the four meta-analyses (in computer-based instruction) the mean effect sizes for published journal articles were higher than those for both unpublished reports and dissertations, but for the other two meta-analyses the mean effect sizes for unpublished reports were higher. However in all four meta-analyses effect sizes were higher in journal articles than in dissertations. Kulik and Kulik urged caution in the interpretation of their results, claiming that the explanation for the relationship of the difference in effect sizes in journal articles and dissertations was 'controversial' (p.272), and not necessarily attributable to publication bias, but more likely to be attributable to the relative inexperience of dissertation writers.

More recent methodological studies of publication bias have been published in the field of healthcare research (e.g., Egger *et al*, 2003). In healthcare research a large methodological effort is put into the topic of publication bias, particularly in the area of reviews of randomized trials. Indeed, the issue has gained such prominence that recently major medical journals have announced they will not publish randomized trials where the protocols have not been registered in advance in a publicly accessible database. This step should, in the long run, prevent small positive trials being published while small negative ones are ignored.

There are examples of meta-analyses in the sphere of educational research where it seems probable that publication bias has affected the magnitude of the pooled effect size. For example, Torgerson *et al* (2003) in a systematic review in the area of adult literacy and numeracy research noted that the field was probably susceptible to publication bias and concluded that small, negative studies evaluating interventions in adult basic education had probably not been published or were probably not in the public domain. Similarly, in their meta-analysis of experimental research into the effectiveness of second-language instruction, Norris and Ortega (2000) discussed the issue of publication bias at length. However, they decided not to include unpublished studies, and cautioned the reader that it was likely that this would lead to 'serious' publication bias (p. 432). Subsequently Truscott (2004) argued that a number of factors had affected the strong positive effect of the review (d = 0.96), including publication bias, and concluded that Norris and Ortega's results were 'substantially inflated' (p.22).

In summary, since 1959 various methodological and empirical researchers have found significant evidence for a file-drawer effect within education and the social sciences. To avoid this bias having a detrimental effect on systematic reviews it is absolutely

essential that, first, the problem is recognised and second, steps are taken to ameliorate this source of bias.

Including unpublished data

Because publication bias has a long and ignoble history, researchers have sought methods both of identifying the problem and minimising its effects.

In their 'practitioner's guide to meta-analysis', Durlak and Lipsey (1991) outline the six major steps involved in conducting an effective meta-analysis and emphasise the procedures critical to the validity of its conclusions. They criticise a common practice in meta-analysis, that of only including published studies on the basis that these will represent the most high quality research in a given field, and suggest that quality criteria should be pre-specified and applied equally to published and unpublished studies (Lipsey and Wilson, 1993).

Researchers can minimise the problem of publication bias by extensive and exhaustive searching, and by including studies that are unpublished but in the public domain. The latter can be achieved by searching the electronic databases that contain unpublished studies: for example, Dissertation Abstracts International; the System for Information on Grey Literature in Europe (SIGLE); and Education Resources Information Center (ERIC).

A justification often given for excluding unpublished studies from a systematic review, even if identified, is as a quality check. For example, for their meta-analysis of controlled trials evaluating systematic phonics instruction versus non-systematic or no

phonics instruction, Ehri and colleagues (2001b) sought only studies from peer-reviewed journals. The authors justified this decision on the basis that unpublished studies were more likely to be of a lower quality than published studies.

Even if we include the grey literature there will still be studies that we simply cannot detect, either because they have not been picked up by even the most sensitive search, or because they do not appear in the databases. If these 'missing' studies have similar characteristics to the identified studies the only issue that their non-inclusion raises is the risk of increasing a Type II error: that is wrongly concluding that there is no statistically significant effect, when in truth there is. But if the missing studies are systematically different from included trials then bias can result. It is important to consider methods of detecting such bias and remedying the situation.

Estimating the extent of publication bias using informal techniques

Once the review has been completed, researchers should attempt to detect publication bias retrospectively and, if found to be present, attempt to correct for it. If publication bias is found to be present in a systematic review, researchers can attempt a sensitivity analysis to assess the potential impact of missing studies on their results. Methods for detecting and assessing publication bias, of the kind described in this section, have a reasonably long history.

Indications of the existence of publication bias can be detected using graphical or statistical methods. Funnel plots are a graphical method first used in educational research (Light and Pillemer, 1984) that can be used to look for patterns in the data. The effect sizes are plotted against the sample sizes (or standard errors) on a graph.

Wang and Bushman (1998) proposed a further graphical method for the detection of possible publication bias: the use of normal quantile plots. They suggest this method because it can difficult to interpret funnel plots by eye whereas it is easier to determine whether or not data fall on a straight line as in the case of normal quantile plots (Wang and Bushman, 1998). The fail-safe *n* test (Rosenthal, 1978; Rosenthal, 1979; Dear and Begg, 1997) is a statistical method to test for possible publication bias. The number of zero studies necessary to reduce the result to non-significance ($p > 0.05$), or to reduce a large effect size to a small effect size, is calculated.

Funnel plot

The simplest and most common method used to detect publication bias is the use of a funnel plot. In a funnel plot the point estimate from each study is plotted against some measure of the precision of the study (usually the standard error or sample size). Those studies with the highest precision will appear high up on the y–axis. A plot with little or no evidence of a publication bias should look like an inverted funnel. The largest study will be at the apex of the funnel with the smaller, less precise studies fanning out in equal measure on both sides of the large study or studies. If publication bias is present we will observe that one side of the funnel is missing (usually the left-hand side indicating negative or null trials missing). Sometimes publication bias might be indicated by a hollowing out of the centre of the funnel plot around the area of no effect. This occurs when statistically significant positive and negative studies are published but those without significant results are not. The results of the funnel plot should then be taken into account when interpreting the conclusions of a systematic review.

For example, a secondary analysis of only the randomized trials included in the National Reading Panel's review of systematic versus non-systematic phonics

instruction (Ehri *et al*, 2001b) noted that there was evidence for possible publication bias, as demonstrated by using a funnel plot (Torgerson, 2003). In this case small negative studies may have been missing. This bias may have led to an over-estimate of the benefits of phonics instruction (Torgerson, 2003), although there were only 13 RCTs in the funnel plot (see below) and asymmetry in a funnel plot is *suggestive* of publication bias.

Limitations of the funnel plot

Asymmetry in a funnel plot may be due to factors other than publication bias. There may be substantive or methodological heterogeneity between the studies. An asymmetrical funnel plot may occur when small, methodologically weak trials produce biased estimates of effect and consequently appear as 'positive' trials when they should be null or negative studies. Sterne *et al* (2001) have outlined five possible reasons for asymmetry in funnel plots: one of these is selection bias, which includes publication bias; the others are true heterogeneity, poor methodological design, heterogeneity due to inadequate outcome measures; and chance. Therefore, asymmetry in a funnel plot is suggestive of publication bias, not proof of it. The possibility of chance accounting for an asymmetrical funnel plot will increase with a declining sample size of trials. Therefore, funnel plots with fewer than 20 trials should be interpreted with caution as an asymmetrical funnel plot may have occurred simply because no trial with a discrepant result has yet been conducted.

Comparisons between published and unpublished studies

Because the results of funnel plots are only suggestive of publication bias they should not really be interpreted in isolation, but rather in conjunction with another method for assessment of possible publication bias. As well as using funnel plots we can also look

at the effect sizes between unpublished and published data. A systematic review of 41 meta-analyses in healthcare research (McAulay *et al*, 2000) found that, in 34% of the cases, removal of the 'grey' literature changed the estimate of effect by 10% or more. In their meta-analysis of adult literacy and numeracy trials, Torgerson *et al* (2003) noted much larger effect sizes among published data (d=0.49, p=0.003) compared with unpublished studies (d=0.26, p=0.13). Similarly, Lipsey and Wilson (1993) also found that unpublished studies had an average effect size somewhat smaller than the average effect size of the published data.

Normal quantile plots

Normal quantile plots can be used to check for possible publication bias. In a normal quantile plot the effect size estimates are plotted against a normal standard distribution with a mean of 0 and a standard deviation of 1 (because a meta-analysis with a large sample size will have a normal distribution) (Wang and Bushman, 1998). If the effect size estimates have a normal standard distribution they will fall approximately along a straight line x = y. If there is a gap in the straight line around zero (where studies with non-significant results are absent, and when the 'true' population effect is zero) or if the line is curved to the right (where non-significant results are missing and when the 'true' effect is different from zero) this will indicate the presence of publication bias. Normal quantile plots can also be used to assess whether or not a set of studies is normally distributed and/or whether they come from a single population or if they are not normally distributed and/or come from two different populations. In the former the data will fall along a straight regression line. In the latter the data will form a 'S' shape, with two clear 'bumps' (Wang and Bushman, 1998).

Fail-safe n test

If there is a suspicion of publication bias then the results of the systematic review can be subjected to the fail-safe *n* test. This is the test that determines the number of studies not retrieved averaging an effect size of zero that would need to exist in order to reduce the summary effect to a non-significant level or to bring the overall probability of a Type I error to a stated level of significance, such as p=0.05 (Rosenthal, 1978; Rosenthal, 1979). Rosenthal (1979) indicated that the findings of a meta-analysis are probably robust if the fail-safe *n* is more than five times the number of reviewed studies plus ten. In their meta-analysis of studies evaluating systematic phonics instruction versus non-systematic or no phonics instruction Ehri *et al* (2001b, p.431) calculated that, for the 43 comparisons they found (in RCTs and CTs) with effect sizes of $d = 0.20$ or greater to be 'statistical exceptions', the existence of 860 comparisons in the unpublished literature with effect sizes below $d = 0.20$ would be required, and they considered this possibility 'unlikely'. However, in his meta-analysis of 11 studies evaluating the effects of reading to young children in schools, Blok (1999) found pooled effect sizes of 0.41 for reading and 0.63 for oral language development. He calculated that the fail-safe *n* for the oral language outcome was 22, i.e. 22 zero effect unpublished studies would be required to reduce this effect size from 0.63 to 0.2. Because this number is not more than five times the number of reviewed studies plus 10 (i.e. 65) it is conceivable that there are sufficient unpublished trials to overturn the result.

Limitations of the fail-safe n test

This method is based on the assumption that the unpublished studies are a random sample of all the studies that were undertaken (Iyengar and Greenhouse, 1988). It combines the results of the selected studies as if they were an unselected sample and then retrospectively assesses the potential effects of publication bias. This assumption

is, of course, unlikely ever to be strictly true. It also over-emphasises the importance of the convention of using p=0.05 to test for statistical significance. It assumes that the average effect of the unpublished studies is in the same direction as the average observed effect (and this may not be the case) so therefore it could be a misleading statistic: the missing unpublished studies could be negative. If this is the case then the calculated fail-safe *n* would be an over-estimate. This method for testing for publication bias has been criticised (see for example, Hsu, 1979 who has suggested a correction in the fail-safe calculation to take into account the possibility that unpublished studies could have negative, rather than null, effects), and Rosenthal himself apparently has not cited it in his later meta-analyses (Evans, 1996).

Conclusions

The problem of publication bias was first recognized in the field of psychology, although many authors have raised the issue of publication bias over the last 40 years in other fields, in particular in healthcare research. Much methodological work has recently been undertaken in this field (see for example Egger *et al*, 2003) and in some areas of the psychological and the social sciences (Sterling *et al,* 1995; Lipsey and Wilson, 1993) to demonstrate the existence of the problem, and to illustrate ways of correcting for it retrospectively.

Sutton *et al* (2000) and Egger *et al* (1997) have demonstrated that, in the field of healthcare research, many meta-analyses do not consider the effect of publication bias on their results. Sutton *et al* analysed 48 systematic reviews in the Cochrane Database of Systematic Reviews. Twenty-three meta-analyses were estimated to have some degree of publication bias (with random effects model). The authors estimated that

about half of meta-analyses may be subject to some level of publication bias and about a fifth have a strong indication of missing trials. This analysis concluded that publication bias was common within the sample of meta-analyses, but that in most cases the bias did not affect the conclusions. The authors deduced that around 5-10% of meta-analyses could have been interpreted incorrectly because of publication bias.

The issue of publication bias is an important threat to evidence-informed research and policy-making. For example, in a systematic review and meta-analysis of studies evaluating adult literacy interventions, an overall benefit of intervention was observed. However, the funnel plot suggested that there was evidence for possible publication bias so the authors concluded that the results should be treated with caution and ideally be confirmed in a large RCT (Torgerson *et al*, 2003).

Publication bias is an important threat to the validity of systematic reviews. Researchers undertaking reviews need to be aware of the problem and investigate the possibility of publication bias in their reviews. Readers of systematic reviews should always be aware that a review containing lots of small positive trials is particularly threatened by possible publication bias. Many of these small trials may have false positive results, and it is possible that small trials containing negative results (whether false or true) have been undertaken but have not been included in the review. There are a number of ways in which the problem can be limited prospectively. In order to attempt to prevent the problem, journal editors should encourage the submission of good quality, but negative or null studies. A recent, even more extreme, suggested development is to have journals dedicated to the publication of null results. Researchers have a responsibility to ensure the timely submission of their trials for publication whatever their results. In the field of healthcare research it has been suggested that a

way of limiting the consequences of publication bias is to set up trial registries in all areas of research and critically assess the process of peer review (Begg and Berlin, 1988). The process of setting up trial registries is well under way for healthcare trials. Such a system would reduce the problem of publication bias in research in education and the social sciences.

An important finding from this overview of the publication bias literature is the relative paucity of methodological work undertaken within the field of educational research. Apart from work by Kulik and Kulik (1989) virtually all the methodological work within the social sciences has been driven by methodologists working in the field of psychology. It is important, therefore, that, given the current emphasis on the use of systematic review methodology in educational policy-making, more methodological research is done in this area. Therefore, a contribution that this item makes to knowledge in educational research is to add significantly to the methods research in publication bias in the educational literature.

In the next chapter I will assess whether systematic reviews in an important educational field, literacy learning in English, are susceptible to publication bias and what steps, if any, the authors have taken to reduce this risk.

Chapter 3: Assessment of publication bias in literacy learning[8]

Introduction

As outlined in the previous chapter, it is widely accepted that systematic reviews and meta-analyses may be subject to bias because non-significant or negative studies are less likely to be published. To limit potential bias due to non-publication, rigorous systematic reviews should try and identify 'grey' or 'fugitive' literature. Methodological research in healthcare systematic reviews has estimated that around one third of meta-analyses contain grey literature (McAulay *et al*, 2000; Moher *et al*, 1999). A tertiary review of publication bias in healthcare research (1993-2003) identified 26 studies that looked for possible publication bias (Dubben and Beck-Bornholdt, 2005) with the presence of publication bias being reported in 23 out of the 26 studies. Interestingly, the authors also looked at the presence of publication bias, in the publication bias literature, and found no evidence for such bias. However, the authors noted that the number of studies was low (26): therefore the power to detect possible publication bias was correspondingly weak.

The extent to which systematic reviews in literacy research attempt to identify fugitive literature is unknown. Also unknown is the extent to which research in this field is susceptible to possible publication bias.

The aims of this chapter are, therefore: to assess the extent to which the 14 systematic reviews in literacy learning identified in the tertiary review (Item 2 of the portfolio) included fugitive studies; to assess the extent of possible publication bias in the field of

[8] A slightly revised version of this chapter has been submitted to *Reading Research Quarterly*

literacy learning by comparing the effect sizes and sample sizes between published and unpublished literature, in order to confirm or refute the widespread belief that unpublished studies tend to have smaller sample sizes than published studies, and that more unpublished studies than published studies have small or insignificant effect sizes; and to apply two graphical techniques (funnel plots and normal quantile plots) to a sample of the 14 systematic reviews in order to see if these tools are useful in identifying the presence of publication bias.

There are a number of ways to graphically assess the potential existence of publication bias. Two relatively simple graphical approaches to ascertain the presence of publication bias are the techniques described in the previous chapter of drawing funnel plots and normal quantile plots. In the tertiary review (see Item 2) the only reviews that explicitly used a graphical method of assessing publication bias, and these used funnel plots, were two reviews that I conducted recently (Torgerson and Elbourne, 2002; Torgerson and Zhu, 2003) and the review conducted by Jeynes and Littell (2000). To the best of my knowledge, the use of normal quantile plots to detect possible publication bias is rare in healthcare research and has not previously been undertaken in the field of educational research.

Methods

To investigate the prevalence of publication bias, the following data were extracted from each of the 14 included reviews: whether or not the review included a search of the 'grey' literature, i.e., whether the inclusion criteria of the review may have exaggerated publication bias (e.g., by only including peer-reviewed papers); whether or not the review included at least one item of 'grey' literature (and the total percentage of 'grey'

literature included); whether or not the review mentioned publication bias; whether the review investigated the possibility of the presence of publication bias through, for example, the drawing of a funnel plot or the calculation of the fail-safe *n* (or any other method), and whether, if publication bias was detected in the review, the reviewers sought to correct for it.

In order to investigate how the 'grey' literature varied according to sample size, an average sample size for the published and unpublished literature for each of the reviews that contained fugitive literature was calculated. The sample sizes within the individual meta-analyses were too small to justify the use of a t-test in order to test for statistical significance individually in the reviews. Also, there was the increased probability of a Type I error as seven meta-analyses were included in this analysis. Therefore, a t-test was performed to test for statistical significance (using the Statistical Package for the Social Sciences) with a larger sample size using the combined mean sample sizes.

The effect sizes of the included reviews were cross-tabulated with information about whether or not the reviews included at least one item of 'grey' literature. This was undertaken in two ways: firstly using the combined effect sizes of all study types; and secondly using the combined effect sizes of only the RCTs.

A sub-sample of the systematic reviews identified in the tertiary review was investigated by drawing funnel plots of all the included reviews that had sufficient RCTs. The smaller the number of included RCTs the greater is the likelihood of asymmetry in the funnel plot being caused by chance. The funnel plots plotted the effect sizes along the x-axis and the sample sizes along the y-axis using the computer software package 'Stata'. An informal examination of all the included funnel plots was

undertaken in order to ascertain how many conformed to the classic funnel shape. As outlined in the previous chapter, funnel plots should not be interpreted in isolation but should be examined in conjunction with other methods. Therefore for each of the reviews where it was possible to draw a funnel plot a normal quantile plot was also drawn in order to compare the results. The normal quantile plots plotted the observed value against the expected normal value, using the statistical package 'Stata'. For a systematic review to be included it had to be possible to distinguish between the randomized controlled trials (RCTs) and controlled trials (CTs) within the review. A maximum of one pooled effect size from each review was included. The mean effect size was the one that used the most homogeneous outcome variable and was deemed to measure the most educationally significant outcome: for interventions in phonological awareness, phonemic awareness or phonics instruction this was reading accuracy or reading comprehension (or mean effect size where this was not available); for writing this was holistic writing quality; for spelling it was the number of words correctly spelled in a list; for all other reading interventions a reading outcome (reading comprehension); and for meta-cognitive interventions a meta-cognitive skills outcome. These decisions were made before I looked at the results. In reviews where there was only one pooled effect size it was automatically selected.

Results

A total of 14 systematic reviews were identified in the tertiary review. Table 3.1 shows all the reviews with data about publication bias.

Table 3.1: 14 reviews in the tertiary review with information about publication bias

Author, date	'Grey' literature searched	Contains at least one item of 'grey' literature?	Publication bias mentioned?	Method for assessing potential publication bias?	If publication bias found was it addressed?
Bangert-Drowns (1993)	Y	Y (65%: 20 out of 31 studies)	N	N	N/A: no method used to investigate potential publication bias
Blok (1999)	Y	Y (40%: 4 out of 10 studies)	Y	Y (fail-safe *n)*	N/A publication bias not thought to be a problem ('...this fail safe number in the present example is a total of 22 studies. In other words, there should be 22 unpublished studies with a zero result in order to reduce the mean effect from 0.63 to 0.20. No matter how exact this number is, it is still difficult to evaluate for lack of insight into the actual number of unpublished studies and their results', p.365)
Bus *et al* (1995)	Y	Y (17%: 5 out of 29 studies)	Y	Y (fail-safe *n*)	N/A: publication bias not thought to be a problem
Bus and van IJzendoorn (1999)	N	N	N	N	N/A: no method used to investigate potential publication bias
Ehri *et al* (2001a)	N	N	N	N	N/A: no method used to investigate potential publication bias
Ehri *et al* (2001b)	N	N	Y	Y (fail-safe *n*)	N/A: (publication bias 'unlikely')
Elbaum *et al* (2000)	Y	Y (50%: 15 out of 30 studies)	N	N	N/A: no method used to investigate potential publication bias
Gersten and Baker (2001)	Y	Y (8%: 1 out of 13 studies)	N	N	N/A: no method used to investigate potential publication bias

Haller *et al* (1988)	Y	N/A*	N	N	N/A: no method used to investigate potential publication bias
Jeynes and Littell (2000)	Y	Y (57%: 8 out of 14 studies)	Y	Y (funnel plot)	N/A :('the funnel plots that emerged demonstrated the kind of pattern one would expect, indicating little or no publication bias that would undermine the results of our meta-analysis', p.31)
Mathes and Fuchs (1994)	Y	Y (30%: 3 out of 10 studies)	N	N	N/A: no method used to investigate potential publication bias
Torgerson and Elbourne (2002)	Y	N	Y	Y (funnel plot)	N/A: (funnel plot 'not suggestive of publication bias')
Torgerson *et al* (2002)	Y	Y (14%: 1 out of 7 studies)	N	N	N/A: no method used to investigate potential publication bias
Torgerson and Zhu (2003)	Y	N	Y	Y (funnel plot)	N/A: funnel plot not suggestive of publication bias.

*N/A = information not available in the review

Detection and correction of publication bias

Out of the 14 included systematic reviews only six reviews mentioned publication bias as a potential source of bias for the review (eight did not), although 11 reviews searched for and included 'grey' literature (three did not). Of these 11 reviews that searched for 'grey' literature, eight included at least one such item. One review (Haller *et al*, 1988) contained insufficient information to ascertain how many of the included studies were published and unpublished, and two reviews searched for grey literature but found none that met the inclusion criteria for the reviews (Torgerson and Elbourne, 2002; Torgerson and Zhu, 2003). Therefore, eight (57%) of the meta-analyses in the tertiary review contained unpublished research. A total of six reviews used an informal or formal method for detecting the possible presence of publication bias (eight did not); and of these none found any evidence for the possibility of publication bias.

In the sub-sample of eight reviews that included 'grey' literature, it accounted for between 8% and 65% of the studies in a meta-analysis. Overall, 'grey' literature accounted for 40% of the studies included in these reviews. Most of the unpublished literature comprised unpublished doctoral theses, but unpublished masters dissertations, technical reports and conference abstracts were also represented.

Did the 'grey' literature vary according to sample size?

In Table 3.2 the average numbers of unpublished and published studies in each review are presented. In this analysis all of the studies from each of the reviews are included. For most of the reviews this is experimental literature (randomized controlled trials and controlled trials). The only exceptions are Bus *et al* (1995) and Mathes and Fuchs (1994), which contain correlational and retrospective studies.

In this table, the average sample sizes between published and unpublished studies are compared. The hypothesis being tested by this table is that the unpublished studies will have smaller sample sizes than the published studies.

Table 3.2: Comparison of average sample sizes: unpublished and published studies

Author, Date	Average n unpublished (standard deviation)	Average n published (standard deviation)	Mean difference (95% CIs)	t-test
Bangert-Drowns (1993)	72.01 (61.02)	69.4 (49.57)	2.67	
Blok (1999)	64.5 (41.52)	112.67	48.17	
***Bus *et al* (1995)**	42 (17.07)	74.48 (94.27)	32.48	
Elbaum *et al* (2000)	38.92 (25.62)	24.69 (17.12)	14.23	
Gersten and Baker (2001)	36	47	-	
Jeynes and Littell (2000)	103.86 (132.67)	146.75 (65.21)	42.89	
Mathes and Fuchs (1994)	57.33 (30.55)	64.00 (34.76)	6.67	
Torgerson *et al* (2002)	16	72.66	-	
Total (excluding Gersten and Baker, 2001 and Torgerson *et al*, 2002)	**62.36 (65.53)**	**68.68 (70.62)**	**6.32 (-31.76 to 19.11)**	**p=0.62 not sign.**

*One extreme outlier removed before calculation of average n.

The data in the table do not support the hypothesis. Pooling all the published and unpublished studies shows no statistically significant differences between the unpublished and published literature (mean difference of 6.32, 95% confidence interval of the difference –31.76 to 19.11, p=0.62).

Did the 'grey' literature vary according to effect size?

One of the meta-analyses (Bus *et al*, 1995) tested for an interaction between publication status and effect size. There were 29 studies in this review: 24 published studies and 5 unpublished studies. A one-tailed t-test found no significant difference between the effect sizes of published and unpublished studies (p=0.48). Therefore the authors

concluded that in this review 'unpublished reports did not yield significantly lower effect sizes than published reports' (Bus *et al*, 1995, p.14).

The sample of reviews was examined to ascertain whether or not the grey literature varied according to effect size. In Tables 3.3 and 3.4 the effect sizes of the 14 systematic reviews are shown by whether or not they included at least one item of 'grey' literature. In the first Table (3.3) all study types and their summary effect sizes are shown, whilst in Table 3.4 the effect sizes restricted to randomized and controlled trials are shown. It has not been possible to include the review by Haller in these tables because there is insufficient information in the review.

Table 3.3: Cross-tabulation of whether or not grey literature was included and effect size: all studies, including correlational, longitudinal, experimental (RCTs, CTs, pre-/post-test) and retrospective studies.

Grey literature included	Grey literature not included
Positive	**Positive**
Bangert-Drowns (1993): **0.27***	Bus and van IJzendoorn (1999): **0.70***
Blok (1999): **0.41***	Ehri *et al* (2001a): **0.53***
Bus *et al* (1995): **0.59***	Ehri *et al* (2001b): **0.41***
Elbaum (2000): **0.41***	Torgerson and Elbourne (2002): **0.37**
Mathes and Fuchs (1994): **0.36***	Torgerson and Zhu (2003) (word-processing): **0.89***[9]
Gersten and Baker (2001): **0.81***	
Jeynes and Littell (2000): **0.65***	
Torgerson *et al* (2002): **0.19**	
Negative	**Negative**
	Torgerson and Zhu (2003) (computer-mediated texts): **– 0.05**

*statistically significant at 95% level

[9] Torgerson and Zhu, 2003 contained three separate meta-analyses. One of these replicated the earlier Torgerson and Elbourne, 2002 review on ICT and spelling and has therefore been excluded. The other two have been included separately.

Table 3.4 Cross-tabulation of whether or not grey literature included and effect size: experimental studies: RCTs (and CTs)

Grey literature included	Grey literature not included
Positive	**Positive**
Bangert-Drowns (1993): **0.31 (0.39*)**	Bus and van IJzendoorn (1999): **0.70***
Elbaum (2000): **0.56* (0.17*)**	Ehri *et al* (2001a): **0.63* (0.40*)**
Gersten and Baker (2001): **1.19* (0.71*)**	Ehri *et al* (2001b): **0.45* (0.43*)**
Torgerson *et al* (2002): **0.19**	Torgerson and Elbourne (2002): **0.37**
	Torgerson and Zhu (2003) (word-processing): **0.89***
Negative	**Negative**
	Torgerson and Zhu (2003 (computer-mediated texts): **–0.05**

*statistically significant at 95% level

Neither table 3.3 nor 3.4 indicates an association, in this sample of meta-analyses, between effect size and statistical significance and whether or not unpublished data were included in the meta-analysis.

All study types

Where all study types were included and where 'grey' literature was included, effect sizes were all positive ranging from 0.19 (small) to 0.81 (large). Seven out of eight of the effect sizes were statistically significant. Where 'grey' literature was not included effect sizes ranged from –0.05 (very small, negative) to 0.89 (large, positive). The negative effect size was not statistically significant; and four out of the five positive effect sizes were significant.

Trials only

Where only trials were included and 'grey' literature was included, again all effect sizes were positive and ranged from 0.19 to 1.19 for RCTs and from 0.17 to 0.71 for CTs. Three out of four of the positive effect sizes were statistically significant. Where 'grey' literature was not included effect sizes ranged from –0.05 to 0.89 for RCTs and 0.40 to

0.43 for CTs. Again, the negative effect size was not statistically significant; and four out of the five positive effect sizes were significant.

These results do not present evidence for publication bias in this set of meta-analyses in literacy research. Rather these results suggest that exclusion of grey literature for meta-analysis does not lead to a tendency to result in a smaller or statistically non-significant average effect of the pooled studies.

These data must be treated with caution as the number of meta-analyses included is relatively small and therefore the power of this analysis to show any significant association is correspondingly low.

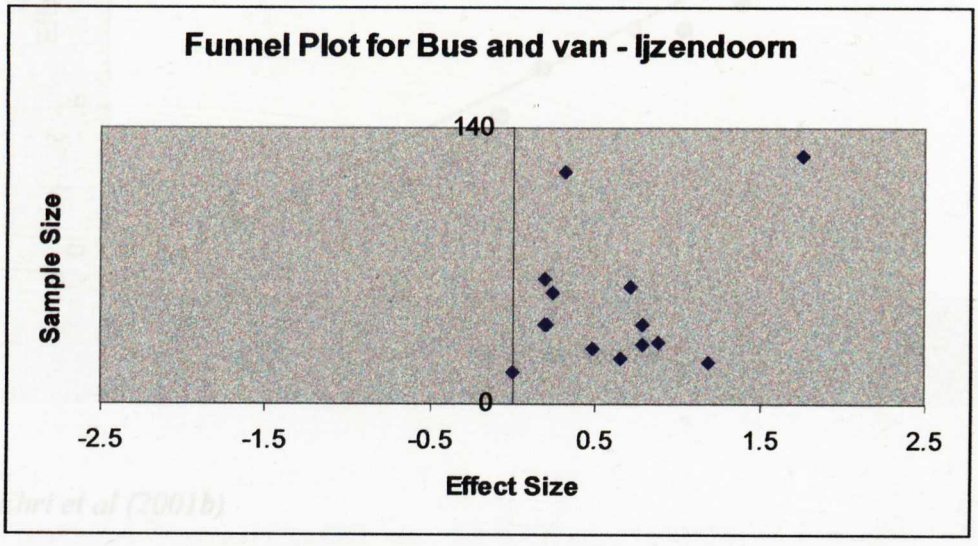
Graphical assessment of publication bias in individual reviews

To assess graphically whether any *individual* systematic review had publication bias, it was necessary to undertake an analysis of each review that had sufficient studies to permit the exploration of publication bias through funnel plots and normal quantile plots. It was only possible to undertake the drawing of funnel plots and normal quantile plots for three of the included studies. The other eleven reviews had to be excluded because they did not give individual effect sizes for included RCTs (seven cases) or because they included too few RCTs (four cases). Therefore funnel and normal quantile plots were drawn for the following three reviews: Bus and van IJzendoorn, (1999); Ehri *et al* (2001b); Torgerson and Elbourne (2002).

Bus and van IJzendoorn (1999)

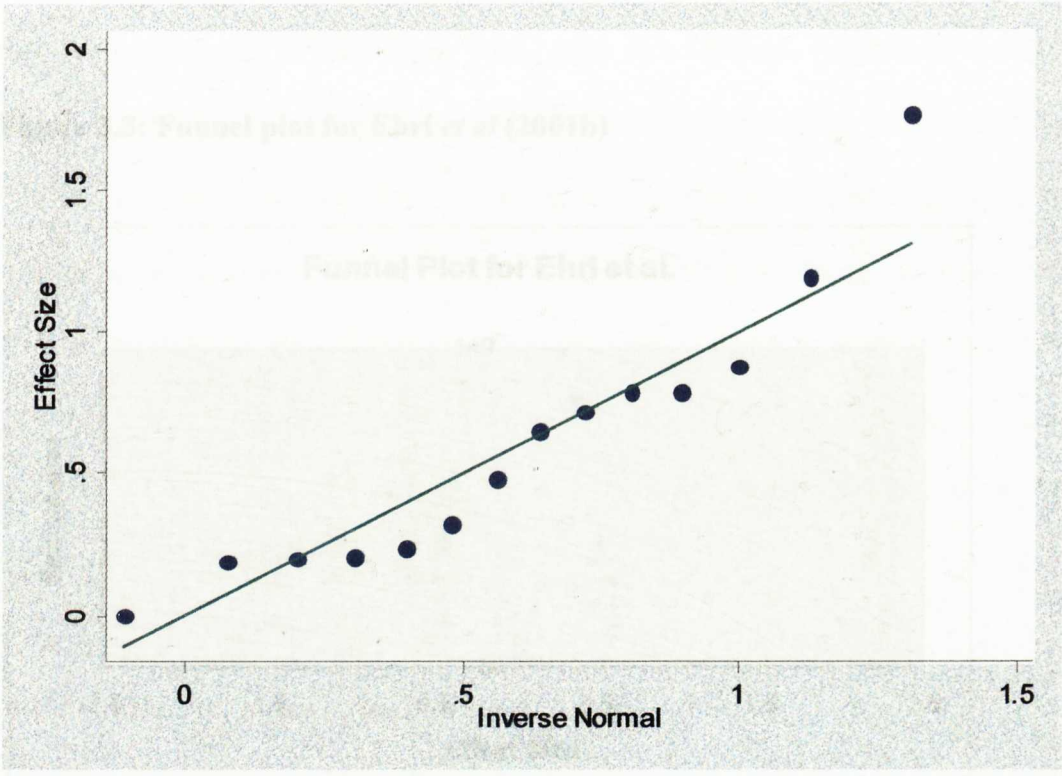
Figure 3.1 shows the funnel plot for Bus and van IJzendoorn (1999). Note there are no trials showing a negative effect, despite relatively small sample sizes. This might be an indicator of publication bias.

Figure 3.1: Funnel plot for Bus and van IJzendoorn (1999)



In Figure 3.2 the normal quantile plot is shown. The plot shows an 'S' shape, which may be suggestive of the trials coming from two or more different populations. Because of this, and because there is a suggestion of publication bias from the funnel plot, it would be sensible to treat any meta-analysis from this study with caution. Indeed, before important policy decisions are made based on this study it may be worth carefully replicating the review in order to check its quality.

Figure 3.2: Normal quantile plot for Bus and van IJzendoorn (1999)



Ehri et al (2001b)

In Figure 3.3 the funnel plot of RCTs included in Ehri *et al* (2001b) is shown. The plot is distinctly asymmetrical, and this could be possible evidence of publication bias. However, because the number of studies is relatively small (i.e., 13), this needs to be interpreted with caution because, by chance, one or two negative, small studies may not yet have been conducted. Nevertheless, in the Ehri *et al* (2001b) review of systematic phonics instruction interventions, only peer-refereed journal articles were included: an inclusion criterion which invites publication bias. As figure 3.3 shows, there were no studies reporting a negative effect of systematic phonics instruction compared with all forms of control despite the small sample sizes of some of the included studies. Although the conclusions of the review suggested that systematic phonics teaching is an effective strategy, this conclusion might have been modified if there had been evidence of possible publication bias. Ideally, to allow a more secure interpretation of the

effectiveness of phonics instruction, all trials, including unpublished material, ought to have been included.

Figure 3.3: Funnel plot for Ehri *et al* (2001b)

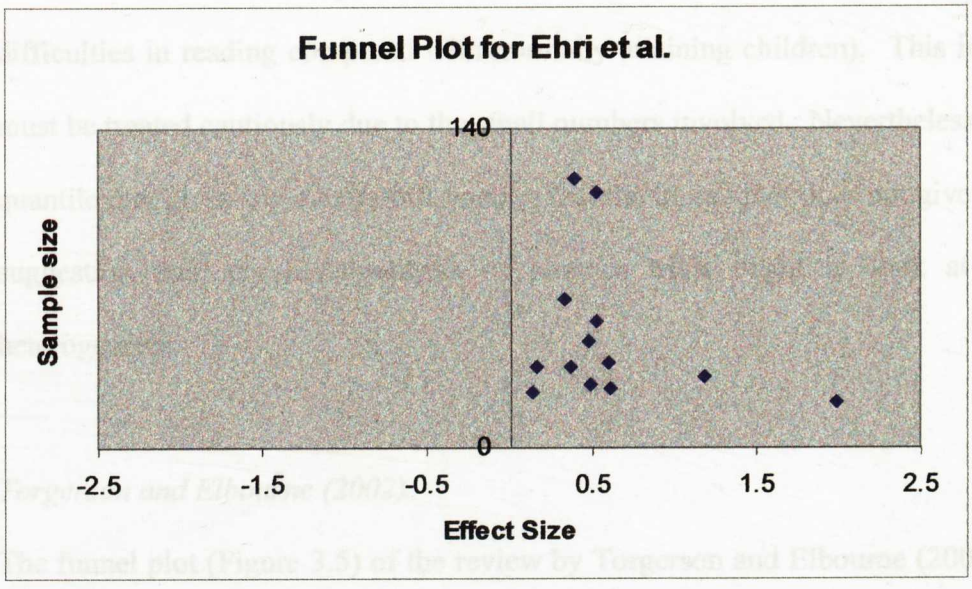


Figure 3.4: Normal quantile plot for Ehri *et al* (2001b)

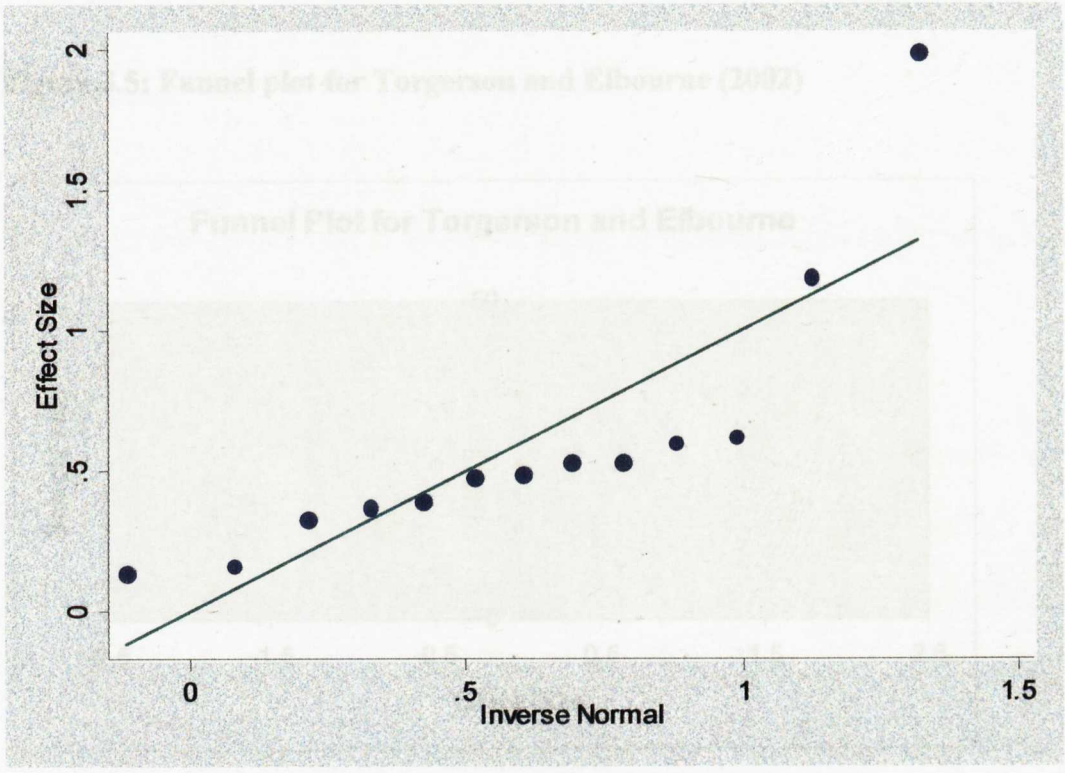
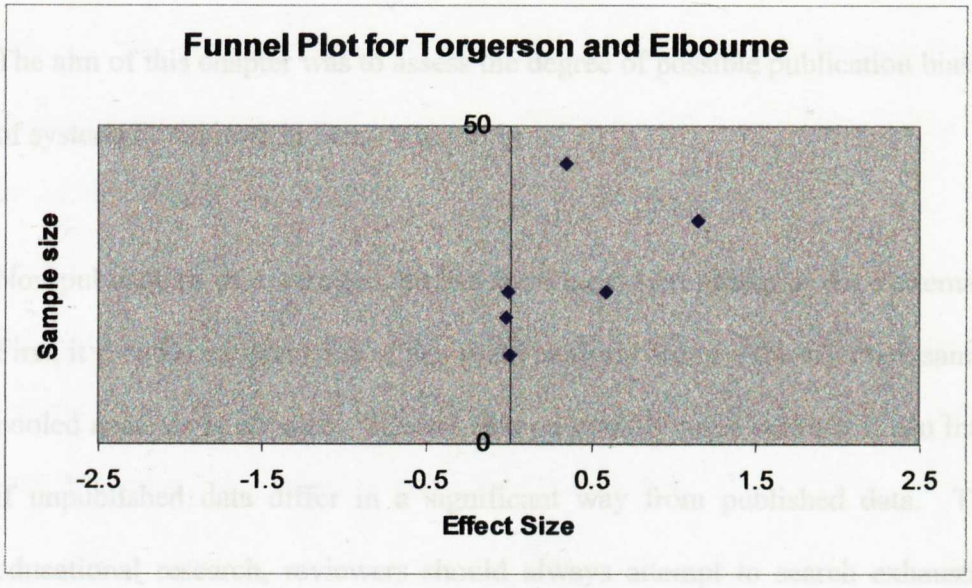


Figure 3.4 shows the normal quantile plot for the Ehri *et al* (2001b) study. In addition to absence of negative or null studies, the plot takes an 'S' shape. This suggests that the data comprising the meta-analysis come from two different population distributions. This might happen if, for example, systematic phonics instruction were highly effective within one group of children but less effective in another population (e.g., children with difficulties in reading compared with normally attaining children). This interpretation must be treated cautiously due to the small numbers involved. Nevertheless, the normal quantile plot gives some extra information that the funnel plot does not give: namely the suggestion that any meta-analysis of phonics trials ought to look at sources of heterogeneity.

Torgerson and Elbourne (2002)

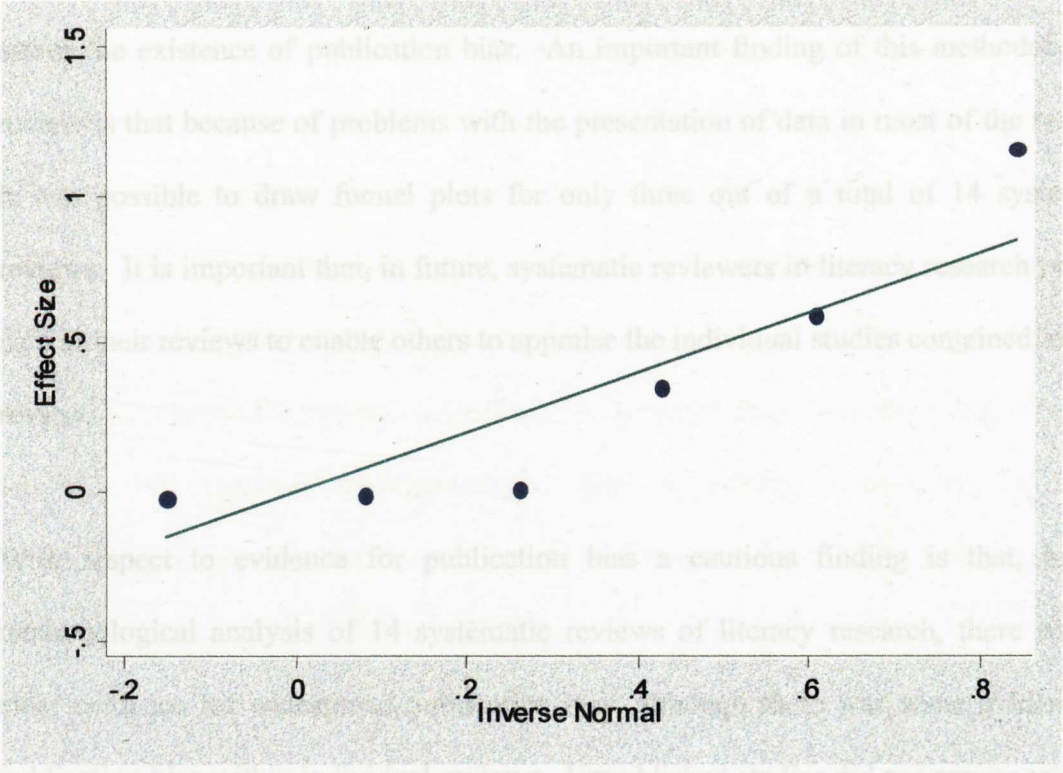
The funnel plot (Figure 3.5) of the review by Torgerson and Elbourne (2002) shows no evidence for publication bias, although given the small number of studies involved (6), it cannot be concluded that publication bias is absent.

Figure 3.5: Funnel plot for Torgerson and Elbourne (2002)



The normal quantile plot (Figure 3.6) is supportive of the funnel plot, in that the identified studies fall closely to the regression line.

Figure 3.6: Normal quantile plot for Torgerson and Elbourne (2002)



Conclusions

The aim of this chapter was to assess the degree of possible publication bias in a sample of systematic reviews in literacy learning.

Non-publication of controlled studies introduces two problems for systematic reviews. First, it reduces the precision of any meta-analysis because the effective sample size of a pooled analysis is reduced. Second, and potentially more serious, it can introduce bias if unpublished data differ in a significant way from published data. Therefore, in educational research, reviewers should always attempt to search exhaustively in the

published and unpublished literature to avoid the effects of publication bias, check for the presence of publication bias (preferably using more then one method) and perform a sensitivity analysis to assess the potential impact of missing studies.

Two graphical techniques, funnel and normal quantile plots, were applied to the data to assess the existence of publication bias. An important finding of this methodological review is that because of problems with the presentation of data in most of the reviews it was possible to draw funnel plots for only three out of a total of 14 systematic reviews. It is important that, in future, systematic reviewers in literacy research present data in their reviews to enable others to appraise the individual studies contained in their review.

With respect to evidence for publication bias a cautious finding is that, in this methodological analysis of 14 systematic reviews of literacy research, there was no clear evidence for widespread publication bias, although there was some evidence of publication bias within individual reviews. Unpublished studies did not appear to differ systematically in their sample or effect sizes. The view that unpublished studies tend to be small with negative or null effect sizes is not supported by the analysis of the tertiary review. On the other hand, it was not possible to include studies that have been undertaken but not recorded anywhere. These studies may, indeed, differ in ways that could bias results.

Policy-makers interpreting the results of meta-analyses in educational research should still be aware of the possible consequences of basing policy decisions on reviews that may be biased due to selective publishing or inclusion.

In this chapter the use of funnel plots was applied to three studies. In two of these studies the funnel plot was suggestive of publication bias, which should indicate that a degree of caution is advisable when using the results of those reviews. This chapter has also explored the use of the normal quantile plot, which to my knowledge is the first time it has been used to assess meta-analyses in educational research. With respect to publication bias in the three reviews where the technique could be used, the normal quantile plots confirmed the interpretation of the funnel plots. Given that funnel plots are notoriously difficult to interpret (Wang and Bushman, 1998), this was a useful exercise. In his recent PhD thesis on formal and informal methods of detecting publication bias, Baldwin cites an inter-rater Kappa statistic of 0.57 (moderate) of agreement between the two raters (Baldwin, S., personal communication, Aug 5th, 2005; Shadish, W., personal communication, Aug. 5th, 2005). Similarly, a recent methodological paper in the healthcare field demonstrated empirically that funnel plots are difficult to interpret (Terrin *et al*, 2005).

The normal quantile plots contributed important information to interpretation of two of the reviews (Ehri *et al*, 2001b; Bus and van IJzendoorn, 1999), which was extra to that obtained from the funnel plots. In both of these studies the normal quantile plot was suggestive of heterogeneity of included studies. This finding is important. If studies with educational heterogeneity are inappropriately placed into a meta-analysis this can be misleading. We may erroneously conclude, for example, that an educational intervention is beneficial for a wider population of learners when, in fact, it is only really effective among a subgroup of learners. The use of this simple graphical technique can signal caution to the policy maker that it may be wise to invest in more research before a policy is widely implemented.

One of the individual reviews included in this tertiary review did have some suggestion of publication bias in terms of included RCTs. The Ehri *et al* (2001b) review of systematic phonics teaching, despite including some very small trials, found no negative or null studies. The associated funnel plot appears to be suggestive of publication bias. Furthermore, the normal quantile plot is suggestive of heterogeneity. Because this review deliberately did not include any unpublished literature but did set out a clear, replicable search strategy and inclusion criteria, in the following chapter the review is replicated to include unpublished data in order to ascertain whether or not inclusion of the grey literature could significantly have altered the original reviewers' conclusions. In addition, as well as possible publication bias adversely affecting the review's conclusions, it is also important to ascertain that the meta-analysis was pooling homogeneous studies, otherwise the summary result may be biased. The review is also updated from the time of the original review (2000) to the present (2005) in order to see if any further published or unpublished studies have been undertaken. This updated review forms the basis of the next chapter.

Chapter 4: Assessment of publication bias within one systematic review: A case study (the phonics review)

Background

One of the reviews retrieved by the tertiary review (Item 2) was a systematic review and meta-analysis of the experimental research (since 1970) evaluating systematic phonics instruction versus unsystematic or no phonics instruction. This review was conducted in the late 1990s by the alphabetic subgroup of the National Reading Panel (NRP) in the USA (Ehri *et al*, 2001b). The aim of the review was to search for, retrieve and synthesise the experimental research base for evidence of the relative effectiveness of systematic phonics instruction, unsystematic phonics instruction or reading instruction without a phonics element. It also wanted to look at the evidence for differential effects depending on different characteristics of learners, for example age or grade level and attainment level (normally attaining children or those experiencing difficulties or disabilities in learning). Ehri *et al* (2001b) only included experimental studies (randomized controlled trials or quasi-experiments) that were published in peer-refereed journals. Potentially relevant trials were identified from electronic searches on the ERIC and PsycINFO databases, and through contact from content experts. Thirty-eight studies met the inclusion criteria for the review and, from these 38 studies, the authors derived 66 effect sizes in order to assess the effectiveness of systematic phonics instruction. The meta-analysis found an overall statistically significant positive effect for phonics instruction on reading of 0.41 (confidence interval (CI) = 0.36-0.47), using a meta-analysis of all trials (randomized trials and quasi-experiments). A meta-analysis of the randomized controlled trials produced a pooled effect size of 0.43. The authors of the review concluded:

> 'Systematic phonics instruction helped children learn to read better than all forms of control group instruction, including whole language. In sum, systematic phonics instruction proved effective and should be implemented as part of literacy programs to teach beginning reading as well as to prevent and remediate reading difficulties' (Ehri *et al*, 2001b, p.393).

The meta-analysis has subsequently been heavily criticised by, for example, Garan (2001a, b, c). In my opinion many of Garan's criticisms are unfounded (for example, including only experimental studies to derive causal inferences; including 'only' 38 studies in the meta-analysis; including only studies with participants with certain learner characteristics) and have been adequately addressed by the authors of the review (Ehri and Stahl, 2001).

More recently, however, the original meta-analysis was replicated (Camilli *et al*, 2003) using the same 38 studies in the original analysis (plus an additional three with phonemic awareness outcomes, minus one study that did not have a 'no treatment' control group). In the re-analysis Camilli *et al* (2003) extracted data from the 40 studies with specific regard to the treatment characteristics, i.e. the 'degree' of phonics or 'mixture' of phonics with other literacy activities (Camilli *et al*, 2003, p.8). They computed effect sizes comparing systematic phonics instruction with the full range of treatment controls (including language based controls) and 'equal study representation' (p.23). Camilli *et al* (2003) found a reduced effect size of d = 0.24 for the comparison between 'systematic' and 'less systematic' phonics instruction, and concluded that 'the advantage of systematic phonics instruction over some phonics instruction is significant but cannot be clearly prioritized over other influences on reading skills' (Camilli *et al*, 2003, p.30). They also used a regression model to show that tutoring and language-based reading activities had similar effect sizes to systematic phonics instruction (d = 0.39 and 0.29 respectively).

In terms of methodology, the original Ehri *et al* (2001b) review had several limitations, which this secondary analysis and update seeks to address. Firstly, its results may have suffered from the effects of publication bias because unpublished trials were excluded. This may have resulted in an over-estimate of effect. Secondly, it included both true experiments (randomized controlled trials) and quasi-experiments (controlled trials). The problem with including both randomized and other controlled trials in a meta-analysis is that pooling two study types can lead to a biased result. Whilst the precision of the estimate may increase (i.e., small confidence intervals), the estimate itself may be incorrect.

Thirdly, a total of 66 comparisons from the 38 trials were included in the review. Double- (and in one case quadruple-) counting of the control groups in comparisons to calculate effect sizes would have had the effect of artificially increasing the sample size and therefore spuriously increasing the precision of the estimated effect (i.e., producing CIs and p values that were too small). The authors of the review acknowledged that by doing this the effect sizes were 'not completely independent across comparisons' (Ehri *et al*, 2001b, p.340). Fourthly, as observed in the previous chapter, there was some indication of heterogeneity between studies, which requires further exploration. Finally, the authors did not quality appraise the individual trials included in the meta-analysis. Including poor quality trials can lead to exaggerated estimates of effect. Ehri *et al* (2001b) did not investigate this possibility in a systematic way.

Publication bias

In the NRP review of systematic phonics instruction interventions, one of the inclusion criteria was that the trials had to be journal articles that had been peer-refereed. Including this criterion could have potentially increased the risk of overestimating the

effect size of the intervention, as it is more likely that negative studies will have been excluded. Ehri *et al* (2001b) did calculate the fail-safe *n* (p.431). However, the calculation was: How many studies of effect sizes below 0.2 (rather than zero or negative estimates) would be required to indicate the 43 comparisons of 0.2 and above were 'statistical exceptions'? As outlined in Chapter 2, the number required with null or negative effects would be fewer. As Figure 3.9 in Chapter 3 of this item shows, there were no studies reporting a negative effect of systematic phonics instruction compared with all forms of control, despite the small sample sizes of the included studies. This figure is suggestive of possible publication bias. Although the conclusions of the NRP review suggest that systematic phonics teaching is probably an effective strategy, this conclusion might have been modified if there had been evidence of publication bias.

The use of randomized controlled trials in effectiveness research

As previously outlined at length in Item 1 of the portfolio, the most robust method of assessing whether an intervention is effective or not is the randomized controlled trial (RCT). This is because, if participants are allocated on any other basis, one cannot be sure whether (except for chance differences) the experimental and control groups were similar before receiving or not receiving the intervention, and it therefore becomes impossible to disentangle the effects of the intervention from the characteristics of the people allocated to it. Techniques can be used to attempt to control for potential confounding from known variables, but they cannot adjust for unknown variables. The two main reasons for using random allocation are to avoid regression to the mean effects and to avoid selection bias. Forming comparison groups using random allocation deals with regression to the mean as it affects both groups equally and the effect is cancelled out in the comparison between the post-test means. Selection bias occurs when the groups formed for comparison have not been created through random

allocation and when the two groups formed are different in some way that can affect outcome.

Methods

The Ehri *et al* (2001b) meta-analysis included 38 true and quasi-experiments. Of these only 13 were randomized controlled trials. Therefore for this review and in order to look for potential publication bias, the grey literature was searched for potentially relevant unpublished trials, and funnel plots of the original and updated reviews were drawn. Only the randomized controlled trials were included from both the original meta-analysis and in the update. Only one effect size was calculated for each study, and this was a mean of the reading accuracy outcomes, as this was the only outcome available for all trials. The individual trials were all quality appraised according to a modification of the CONSORT criteria for assessing the quality of reporting of randomized trials. (The CONSORT criteria are discussed in detail in Chapter 5 of Item 1, and outlined again in Chapters 6 and 7 of this item).

Locating the trials

For the purposes of this update and secondary analysis, searches for unpublished trials (1970 – 2005) and for published and unpublished trials reported between 2000-5 were undertaken. Some of the trials for potential inclusion in this review were identified from the original NRP review (13 randomized controlled trials identified in Ehri *et al*, 2001b). In order to locate any further potentially relevant published or unpublished randomized controlled trials the original searches carried out by Ehri *et al* (2001b) were updated, using searches written specifically for this review but based on the original search terms (Ehri *et al*, 2001b, p.399). ERIC and PsycINFO were searched in this way

for the period 2000 to 2005. For PsycINFO a search strategy was created using the three sets of search terms in the Ehri *et al* (2001b) paper (p.399) combined using 'AND'. This strategy was run for the period 1970-2005. This retrieved 1079 records for the period 1970-2000 and 398 records for the period 2001-5. For the period 1970-2000 the database was sorted by publication type and then only unpublished records were included (103). For the ERIC database the three sets of terms in the Ehri *et al* (2001b) review (p. 399) were combined: 'set 1 AND (set 2 OR set 3)' and run for the period 1970-2000. The database was sorted by publication type and then only unpublished records were included (4462).

A new search was written (based on Ehri *et al*'s search terms) and run on SIGLE (the groups of search terms were combined: 'set 1 AND (set 2 OR set 3)') for the period 1970-present. The results from the searches were imported into EndNote and de-duplicated.

Screening and quality assurance procedures

All of these studies were double screened using titles and abstracts, where available, and on the basis of criteria adapted from the original criteria (Ehri *et al*, 2001b, p.400). Screening was undertaken by two reviewers (including for both databases by the candidate) working independently and then meeting to discuss any differences in decision to include or exclude articles, with the exception of the records retrieved through the re-run of the original search on ERIC (1970-2000). I screened this database, and a random sample of 10% was generated and double screened by a second reviewer. A Cohen's Kappa statistic was calculated to assess the inter-rater reliability of the screening.

Inclusion criteria

Trials with the following characteristics were included: randomized controlled trials focusing on the teaching of phonics in English, and either comparing the effectiveness of instruction in systematic phonics with that of instruction providing unsystematic phonics instruction or no phonics instruction, but where the control condition included reading instruction and where there was sufficient information in the original paper about the details of the control condition (this was not always the case in the Ehri *et al*, 2001b review, as described by Camilli *et al*, 2003). Trials also had to measure reading as an outcome and report statistics permitting the calculation or estimation of effect sizes. They also had to involve interventions that might be found in schools.

Exclusion criteria

Trials were excluded if they were not randomized controlled trials, if they did not evaluate the relative effectiveness of systematic phonics instruction versus no phonics instruction, if they were 'short-term laboratory studies with a limited focus' (Ehri *et al*, 2001b), or if they lacked reading as an outcome or statistics allowing calculation or estimation of effect sizes. All trials that primarily investigated phonemic awareness instruction or phonological awareness instruction were excluded (as in the original review). All trials that investigated the relative effectiveness of different kinds of phonics instruction (e.g., synthetic versus analytic phonics instruction) were excluded, as in the original meta-analysis. In addition, all trials that compared two or more kinds of synthetic phonics instruction were also excluded.

Data extraction and quality appraisal

Data were extracted from each included RCT (Table 4.2) in the following categories: bibliographic details; study design; participants (including specific learner

characteristics); details of the interventions and control group treatments; outcome measures, sample size and effect size. In addition, each of the trials was assessed for quality using an adaptation of the CONSORT statement. Items for quality appraisal included: whether or not allocation was concealed; whether or not there was 'blinded' assessment of outcome, etc.

Calculation of effect sizes

For the trials included in the original review, the average effect size as reported in Ehri *et al* (2001b) was noted. However, for all the trials included from the original review, from the update and from the unpublished literature, an effect size was calculated based on a mean of reading accuracy. One of the main criticisms of the Ehri et al (2001b) review (Garan, 2001a, b, c) was that only a few of the included studies used reading comprehension as an outcome measure. Whilst this is true, it is a limitation of some of the included studies rather than a limitation of the review itself. Only six out of the fourteen included RCTs used reading comprehension as an outcome measure, eight out of the fourteen trials used a spelling outcome measure, and all fourteen used at least one measure of reading accuracy. Where possible standardised test results were used; 'experimenter-devised tests' were only used where there was no alternative standardised test. The comparators for the calculation of effect sizes were interventions using systematic phonics instruction (any kind) compared with control groups using unsystematic or no phonics instruction, but using some kind of systematic reading instruction (e.g., whole word or whole language).

Two of the included RCTs were cluster RCTs (Berninger *et al*, 2003; Brown and Felton, 1993). Because taking the raw Ns of participants in such RCTs tends to give them undue weight in meta-analyses, I needed to calculate an effective sample size after

adjusting for the effects of clustering. I applied the formula: 1+(m-1) x ICC, where m is the average size of the cluster and ICC is the intra-cluster correlation. I used the ICC from a recent RCT of information and communication technology and spelling/reading undertaken with colleagues from York and Sheffield (Brooks *et al*, 2005). This ICC was 0.45. I applied the formula s.d. = $\sqrt{n}$ x SE to calculate the standard deviation in the one paper where the s.d. was not available, but where the standard error (SE) was available (Lovett *et al*, 1989).

Meta-analysis

The main meta-analysis pooled the effect sizes of the individually randomized trials, using the computer software package Stata. To investigate possible sources of heterogeneity, sub-group analyses were performed according to learner characteristics and methodological variation in the trials.

Funnel and normal quantile plots

To investigate the potential for publication bias in the updated meta-analysis I drew a funnel and normal quantile plot. I also explored the relationship between sample size and effect size using a bubble plot.

Quality assurance

For quality assurance purposes, all data extraction (including quality appraisal) of the included studies was undertaken by the candidate and two other reviewers. The reading instruction interventions (and control treatments) and outcome measures were categorised by two reviewers.

Results

Searching and screening

The results of the searching and first and second stages of screening are presented in Table 4.1. De-duplication of the results from the electronic searches at the searching stage was done hierarchically in the order presented in the table, i.e. starting with Ehri *et al* and PsycINFO and then moving down the table to ERIC and SIGLE. Therefore if a paper was identified through searching a database lower down the hierarchy (for example, SIGLE) that had already been retrieved through searching another database higher up the Table (for example ERIC) this record is not shown in the figures. The ordering of the hierarchy is based on past experience of searching for and retrieving RCTs in educational research through electronic databases (Torgerson, 2003).

Table 4.1: Results of first and second stage screening

Electronic database or method of retrieval	Initial 'hits' after de-duplication	No. included at first stage	Not received	Included in update
Ehri *et al* (2001b)	13	13	0	9
PsycInfo 1970-2000	103	1	0	0
PsycInfo 2001-2005	398	19	0	1
ERIC 1970-2000	4462	37	1	1
ERIC 2001-2005	652	14	0	1
SIGLE	61	0	0	0
Contact	2	2	0	2
Total	**5691**	**86**	**1**	**14 RCTs**

A total of 5691 potentially relevant studies were identified through the searching of the electronic databases and through searching the original review. After screening at first stage, 86 potentially relevant papers were identified. Those not already in my possession were sent for through inter-library lending, and re-screened on the basis of the full papers, and using the inclusion/exclusion criteria.

After the second stage of screening a total of 14 RCTs were included in the update. These were: 9 RCTs from the original Ehri review (Brown and Felton, 1990; Greaney *et al*, 1997; Haskell *et al*, 1992; Leach and Siddall, 1990; Lovett *et al*, 1989; Lovett *et al*, 1990; Martinussen *et al*, 1998; Torgesen *et al*, 1999; Umbach *et al*, 1989); one RCT from the updated PsycINFO search (O'Connor and Padeliadu, 2000); one RCT from the search of the ERIC unpublished literature (Skailand, 1971); one RCT from the updated ERIC search (Berninger, 2003), and two RCTs through contact (Torgesen *et al*, 2001; Johnston and Watson, 2004, Exp. 2). One trial was unobtainable or not received by the cut-off date of July 31st 2005. Despite exhaustive searching of the 'grey' literature, only one of the included RCTs was unpublished (Skailand, 1971).

One of the trials originally included in the Ehri *et al* (2001b) meta-analysis (Gittelman and Feingold, 1983) was excluded from my analysis because the trial did not contain a phonics instruction intervention group. It is puzzling that Ehri *et al* (2001b) included this study in the original review. Although it states that one of the interventions was 'motivated reading remediation...following the principles of the phonics method' (Gittelman and Feingold, 1983, p.170), it also states that 'wherever possible, whole word recognition was introduced to enable the development of smooth, efficient, rapid reading and to avoid over-reliance on phonetic word analysis'. Clearly this intervention is *not* systematic phonics instruction. Indeed it closely resembles some of the unsystematic phonics instruction or no phonics instruction conditions used in the Ehri *et al* (2001b) analysis as comparators to systematic phonics instruction. A second trial from the original Ehri *et al* (2001b) review was excluded (Mantzicopoulos *et al*, 1992) because the control condition was not an appropriate comparison as the children did not receive a reading intervention: 'TEACH does not provide direct reading instruction to vulnerable readers' (p.574). Again, Ehri *et al*'s decision to include this study is

puzzling. In addition to the inappropriate control group, this trial suffered from huge attrition. A total of 437 'at risk' kindergarten children were randomized (p.575), but 'only 168 children with complete scores were still in the intervention study at the end of second grade' (p.576), an attrition rate of 269 or 62%. However, the authors claim an attrition rate of 280 (p.582) and in the results table (Table 4, p.582) total n = 87. Clearly this study should have been excluded on two grounds: inappropriate control and huge attrition leading to likely attrition bias. The authors discuss these problems at length in the paper (p.582). Two further trials were excluded because the experimental treatments were varieties of systematic phonics instruction, and the control groups did not receive any comparable reading instruction (Lovett and Steinbach, 1997; Lovett *et al*, 2000).

Also at the second stage of screening, four studies retrieved through the update were excluded because they only compared different types of phonics instruction (Fayne and Bryant, 1981; Sullivan, 1971; Walton *et al*, 2001, Exp. 1; Walton *et al*, 2001, Exp. 2). Three studies were excluded because they compared differences *within* synthetic phonics (Oudeans, 2003; Hatcher *et al*, 2004; Jenkins *et al*, 2004). Four studies were excluded because they did not contain an appropriate control group according to the inclusion criteria (Herrera *et al*, 1997; Vadasy *et al*, 2000; Walton and Walton, 2002; Blachman *et al*, 2004).

It was not possible to ascertain how many children were in each of the intervention and control groups in two of the included papers (Lovett *et al*, 1989; Lovett *et al*, 1990). I therefore wrote to Lovett to ask for the numbers. I also wrote to Ehri and Camilli to ask what they did in their respective meta-analyses. One of the co-authors of the paper, Steinbach (personal communication) informed me that in the Lovett *et al* (1989) paper

there were 60 pupils in the DS intervention and 61 pupils in the OWLS group. Therefore we used these numbers in my calculations. All three respondents stated that in the Lovett *et al* (1990) study there were 18 pupils in each group. Therefore I also used these numbers in my calculations.

Table 4.2 contains information about each of the included RCTs that compared systematic phonics instruction with another form of reading intervention (whole language or whole word reading instruction). The table includes information about study design, participants, intervention and control treatments and effect sizes. Table 4.3 contains the quality assessments of all the included trials. This table is based on the modified CONSORT guidelines for quality assessment of RCTs, and includes assessment of whether the individual trials reported method of random allocation and sample size justification, and whether or not assessment of outcomes was 'blind'.

Table 4.2: Characteristics of the included RCTs.

Author (date)	Study design	Participants	Intervention control	Sample size	Effect size, as calculated by Ehri *et al* (mean of all reported outcomes)	Outcome measures used in calculation of effect size, and notes.	Effect size, as calculated by CJT (mean of word recognition and word attack measures)
Berninger *et al* (2003)	Cluster RCT 24 clusters; 2 children in each cluster	Second grade children at risk for persistent reading problems and disabilities	Word recognition versus reading comprehension (whole language)	48 (word recog. n = 24; reading comp. n = 24) Effective sample size: 34	N/A	WRM word identification and word attack subtests administered; only word attack results reported, because this test showed a positive result: possible researcher bias.	0.3 (-0.38 to 0.97)
Brown and Felton (1990)	Cluster RCT 6 clusters, 3 in each arm (8 children in each cluster)	Children at risk for reading disability in G1	Code emphasis versus context emphasis instruction for acquisition of word identification and decoding skills (synthetic phonics versus look-and-say).	47 (code n = 23; context n = 24) Effective sample size: 12	0.48	WRM word identification and word attack subtests	0.24 (-0.89 to 1.37)
Greaney *et al* (1997)	Ind. RCT	'Disabled readers' G3 – G6	Rime analogy training or item-specific training (onset-rime versus look-and-say)	36 (18 in each group)	0.37	Burt NZ raw score Neale (1988) raw score	0.29 (-0.37 to 0.95)
Haskell *et al* (1992)	Ind. RCT	Normally attaining first grade pupils	Phoneme level training group versus whole-word level training group	24 (12 in each group)	0.14	Experimenter-devised tests: Reading regular	0.07 (-0.73 to 0.87)

						words Reading exception words	
Johnston and Watson (2004), Exp. 2	Ind. RCT	Normally attaining primary 1 children	Synthetic phonics group versus no-letter training group (look-and-say)	92	N/A	BASWRT	0.96 (0.42 to 1.50)
Leach and Siddall (1990)	Ind. RCT	Normally attaining first grade pupils	Direct instruction versus paired reading (look-and-say)	20 (10 in each group)	1.99	Neale (1988) accuracy	0.80 (-0.11 to 1.71)
Lovett *et al* (1989)	Ind. RCT	'Disabled readers', mean age 10.8 years	Decoding skills programme group (DS) versus oral and written language stimulation group (OWLS, whole language)	121 (DS n = 60, OWLS n = 61)	0.39	WRAT-R reading subtest PIAT reading recognition SORT GORT connected text	0.22 (-0.14 to 0.57)
Lovett *et al* (1990)	Ind. RCT	'Disabled readers', mean age 8.4 years	REG ≠ EXC (regular words not taught using the exception words method - phonics approach) versus REG = EXC (regular words taught the exception words method - look-and-say approach)	36 (18 in each group)	0.16	WRAT-R reading subtest GORT accuracy Experimenter-devised test administered but results only given for experimenter-devised tests	-0.19 (-0.85 to 0.46)
Martinussen and Kirby (1998)	Ind. RCT	Kindergarten pupils assessed as low performers on phonological processing	Successive phonological group versus meaning group (whole language)	28 (13 in phonics group; 15 in meaning group). Attrition n = 2 from phonics group	0.62	WRM word attack; word identification; word reading (Ball and Blachman); results at floor for	0.44 (-0.31 to 1.19)

		measures				meaning group in word attack test, therefore not calculated	
O'Connor and Padeliadu (2000)	Ind. RCT	G1 children nominated as 'very poor readers'	Blending versus whole word conditions	12 (6 in each group)	N/A	Experimenter-devised tests – total words read (taught and transfer words)	0.53 (-0.62 to 1.68)
Skailand (1971)	Ind. RCT	Normally-attaining kindergarten children	Grapheme/phoneme group versus whole word (look-and-say) group	42	N/A	Experimenter-devised tests – recall of words; transfer to similar words and syllables	-0.17 (-0.78 to 0.44)
Torgesen *et al* (1999)	Ind. RCT	Kindergarten children with weak phonological skills	PASP versus RCS	90 (45 in each group)	0.33	WRM-R Word attack Word identification	0.07 (-0.34 to 0.48)
Torgesen (2001)	Ind. RCT	Children between the ages of 8 and 10 identified as 'learning disabled'	Embedded phonics versus Auditory Discrimination in Depth Program	50	N/A	WRMWIT-R GORT-III accuracy TOWRE/SWE	-0.31 (-0.87 to 0.45)
Umbach *et al* (1989)	Ind. RCT	First grade students having difficulty with reading	Reading mastery (direct instruction) versus Houghton-Mifflin (look-and-say)	31 (15 in direct instruction, 16 in basal program)	1.19	WRM Word identification Total reading	2.69 (1.72 to 3.67)

Abbreviations (reading outcome measures)

BASWRT = British Ability Scales Word Reading Test

Burt NZ = Burt Word Reading Test, New Zealand Revision

GORT = Gray Oral Reading Test

GORT-JII = Gray Oral Reading Test, 3rd edition

Neale = Neale Analysis of Reading Ability

PIAT = Peabody Individual Achievement Tests

PIAT-R = Peabody Individual Achievement Tests, revised

SORT = Slosson Oral Reading Test

TOWRE/SWE = Test of Word Reading Efficiency, Sight Word Efficiency subtest

WRAT-R = Wide Range Achievement Tests, Revised

WRAT-3 = Wide Range Achievement Tests, 3rd edition

WRM = Woodcock Reading Mastery

WRM-R = Woodcock Reading Mastery, revised

Table 4.3: Quality assessment of included RCTs.

Author, date	Reporting of method of allocation	Sample size justification	Intention to teach analysis	Blinded assessment of outcome	Comments
Berninger *et al* (2003)	N/S	N/S	N/S	N/S	Attrition N/S
Brown and Felton (1990)	N/S	N/S	N	N/S	48 children randomized, yet only 47 mentioned in results section (1 lost from code group).
Greaney *et al* (1997)	N/S	N/S	Y	N/S	No attrition.
Haskell *et al* (1992)	N/S	N/S	Y	N/S	
Johnston and Watson (2004), Exp. 2	N/S	N/S	N	N/S	Attrition n = 7 Random allocation only confirmed through contact with author.
Leach and Siddall (1990)	N/S	N/S	Y	N/S	
Lovett *et al* (1989)	N/S	N/S	Y (for first battery of tests)	N/S	Numbers in each of the treatment groups requested and received from authors. Numbers only available for first battery of tests.
Lovett *et al* (1990)	N/S	N/S	Y	N/S	Numbers in each of the treatment groups requested

					and received from authors.
Martinussen and Kirby (1998)	N/S	N/S	N	N/S	Attrition n = 2 in phonics group. Results at floor for word attack test (meaning group).
O'Connor and Padeliadu (2000)	N/S	N/S	Y	N/S	
Skailand (1971)	N/S	N/S	Y	N/S	
Torgesen *et al* (1999)	N/S	N/S	Y	N/S	
Torgesen *et al* (2001)	N/S	N/S	N	N/S	Attrition n = 10
Umbach *et al* (1989)	N/S	N/S	Y	N/S	

In 11 of the included trials the effect size was positive, and ranged from extremely small (Haskell *et al*, 1992; Torgesen *et al*, 1999), through moderate (Berninger *et al*, 2003; Brown and Felton, 1990; Lovett *et al*, 1989; Martinussen and Kirby, 1998; O'Connor and Padeliadu, 2000), to large (Leach and Siddall, 1990) or extremely large (Johnston and Watson, 2004; Umbach *et al*, 1989). Only the extremely large trials were statistically significant. In three of the included studies the effect size was negative and small (Lovett *et al*, 1990; Skailand, 1971; Torgesen *et al*, 2001), but in no case was this statistically significant. Where a calculated effect size could be compared with the Ehri *et al*, (2001b) mean effect size, in all cases my calculated effect size was smaller. This was probably due to the fact that Ehri and colleagues compared systematic phonics instruction to no phonics or unsystematic phonics instruction controls, but in some cases this involved no reading instruction. In most cases the direction and magnitude of effect was the same. The exceptions to this were Lovett *et al* (1990) and Torgesen *et al* (1999).

Quality assurance

Screening of searches on electronic databases: For databases where two reviewers screened the entire database, the agreement between reviewers was high.

Disagreements occurred only on whether or not the trials should be included according to the intervention criterion. On reviewer was consistently more inclusive (JH), and included in some cases trials that evaluated phonemic awareness instruction or phonological awareness instruction. In all cases agreement to include or exclude was secured after discussion to resolve any differences. For the screening of the 10% random sample of the ERIC database of unpublished literature the Cohen's Kappa measure of agreement was 1 (perfect agreement). Therefore it was not considered necessary for any further double screening to be undertaken.

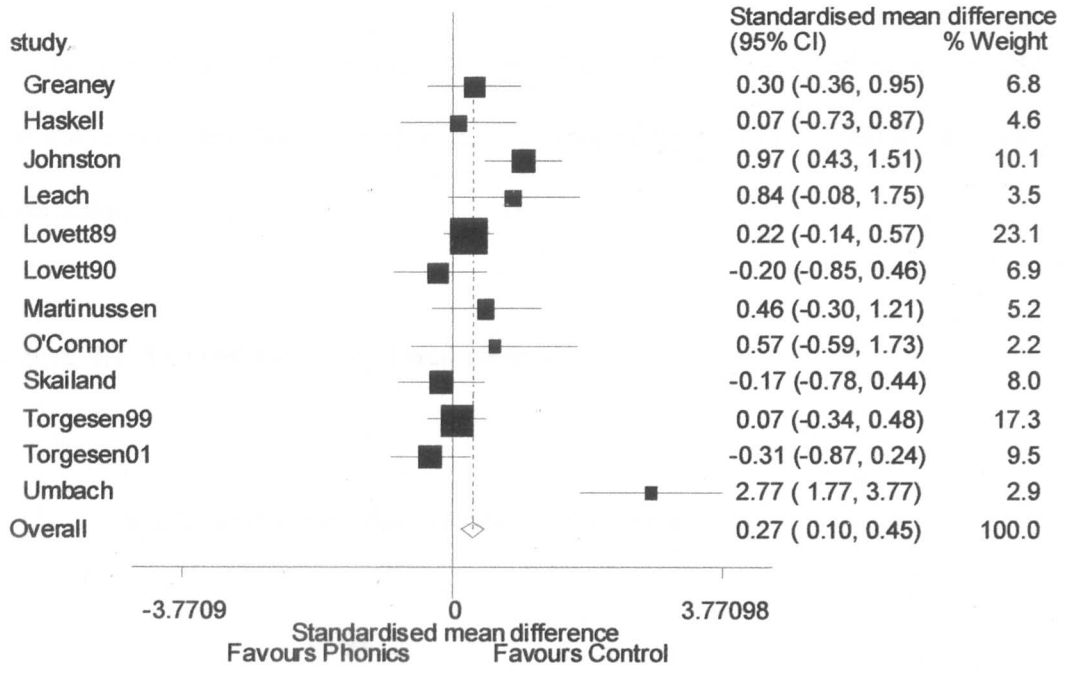
Screening at second stage: Full agreement was established on whether or not to include at second stage of screening (screening of full papers), and the appropriate comparison and outcome measures to be used in the calculation of effect sizes.

Data extraction; quality appraisal; calculation of effect sizes: Initial agreement between the two independent extractions and calculations was high; full agreement was established through discussion.

Meta-analysis of 12 RCTs (main analysis)

Of the 14 RCTs, 12 were individually randomized studies. These were pooled in a meta-analysis (Figure 4.1). The analysis used the standardised mean difference (SMD) (difference between the two means divided by pooled standard deviation) and assumed a fixed effects model, as this was the model adopted by Ehri et al in their meta-analysis (2001b). To ascertain whether there was any difference using an alternative approach to meta-analysis a random effects meta-analysis was also undertaken.

Figure 4.1: Meta-analysis of individually randomized trials



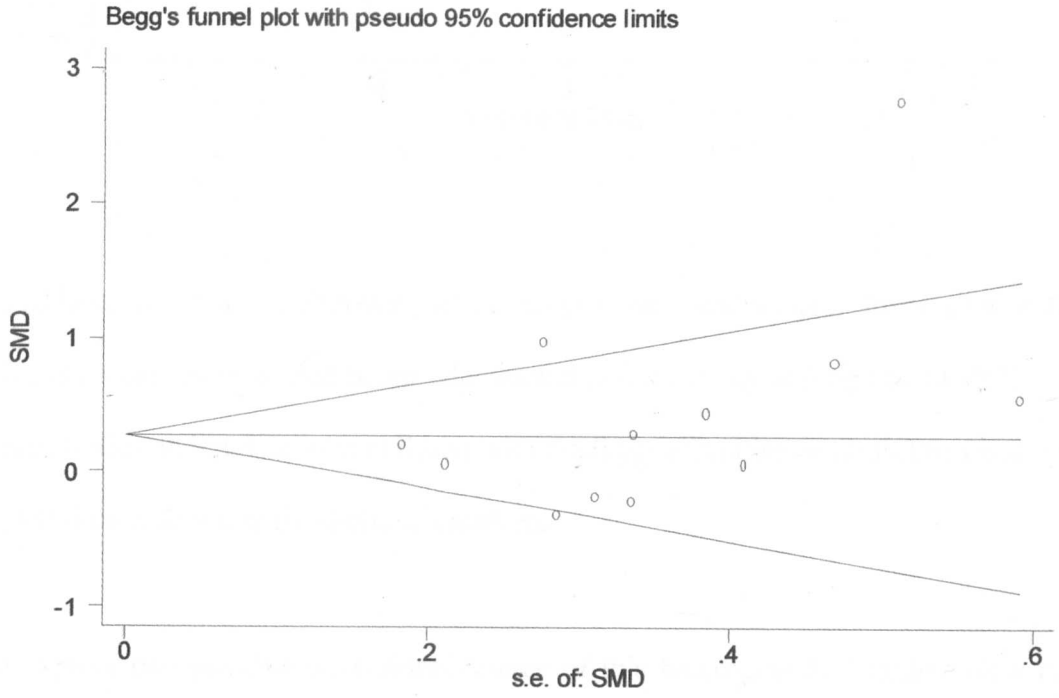
The figure shows that there is a statistically significant effect of phonics instruction on reading accuracy (p=0.002), SMD = 0.27 (0.10 to 0.45). However, the studies displayed significant heterogeneity (Heterogeneity chi-squared = 41.74 (d.f. = 11), p<0.0001). Using the DerSimonian-Laird pooled SMD gave a pooled effect size of 0.38 (95% CI 0.03 to 0.73, p = 0.035).

Publication bias

The updated review only found one unpublished study, with an effect size of -0.17 (a negative result). Publication bias may still be present, however. The average effect size of the revised meta-analysis was 0.27 (95% CI 0.10 to 0.45). For a study to have 80% power to observe this estimate with a 5% significance would require a sample size of approximately 400. Of the studies in the review, all are insufficiently powered to show this difference. Indeed, the average size of the studies included in the review would only have 80% power to observe an effect size of 0.85. This suggests, therefore, there

are similarly powered studies that have smaller, not statistically significant, effect sizes that remain unpublished even within the grey literature. To informally test for potential publication bias in the updated review a funnel plot was drawn and the Egger statistical test for asymmetry was calculated (Egger *et al,* 1997). The resulting funnel plot does suggest asymmetry, but the Egger test for asymmetry is 0.17, which is not statistically significant.

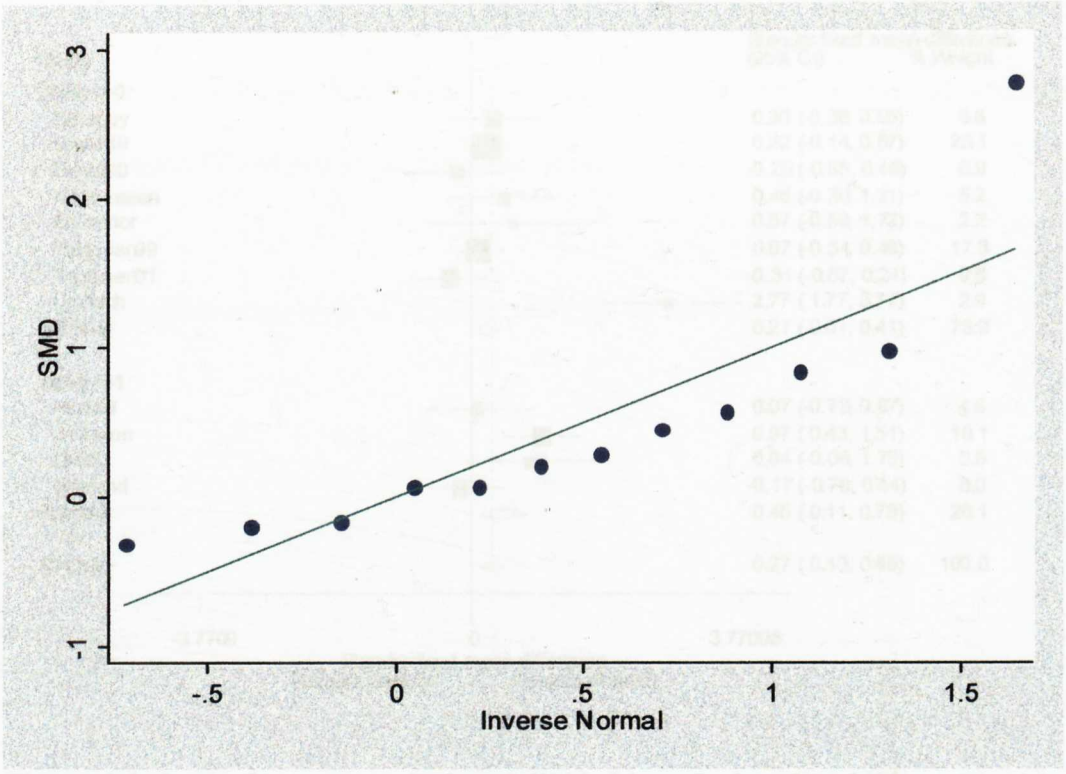
Figure 4.2: Funnel plot of updated review



Study heterogeneity

There was significant heterogeneity in the pooled data ('Q' statistic 46.30, $p<0.001$). In addition, the normal quantile plot (Figure 4.3) is also suggestive of at least two study populations.

Figure 4.3 Normal quantile plot of updated review



To address this issue the following meta-analyses were undertaken. Inspection of Table 4.2 shows that there is significant educational heterogeneity among the 14 RCTs, with some studies undertaken with children with reading difficulties or disabilities and others undertaken with normally attaining children.

To explore one possible educational source of this heterogeneity I undertook a meta-regression to assess whether or not there was an interaction between the effect of phonics instruction and learner characteristics.

Figure 4.4: Meta-analysis subdivided by learner characteristics

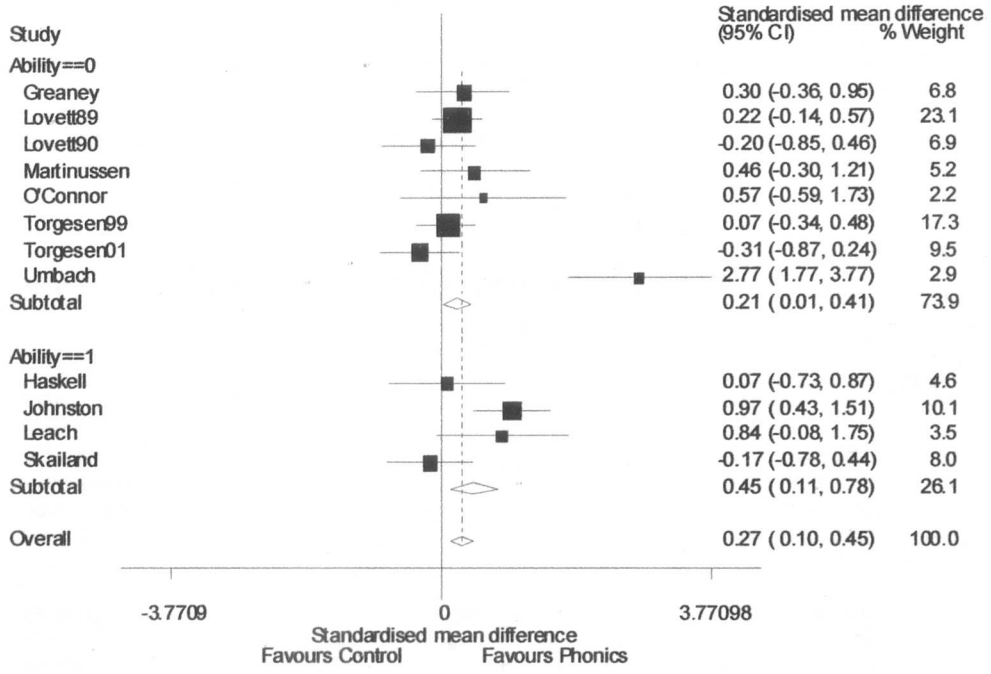
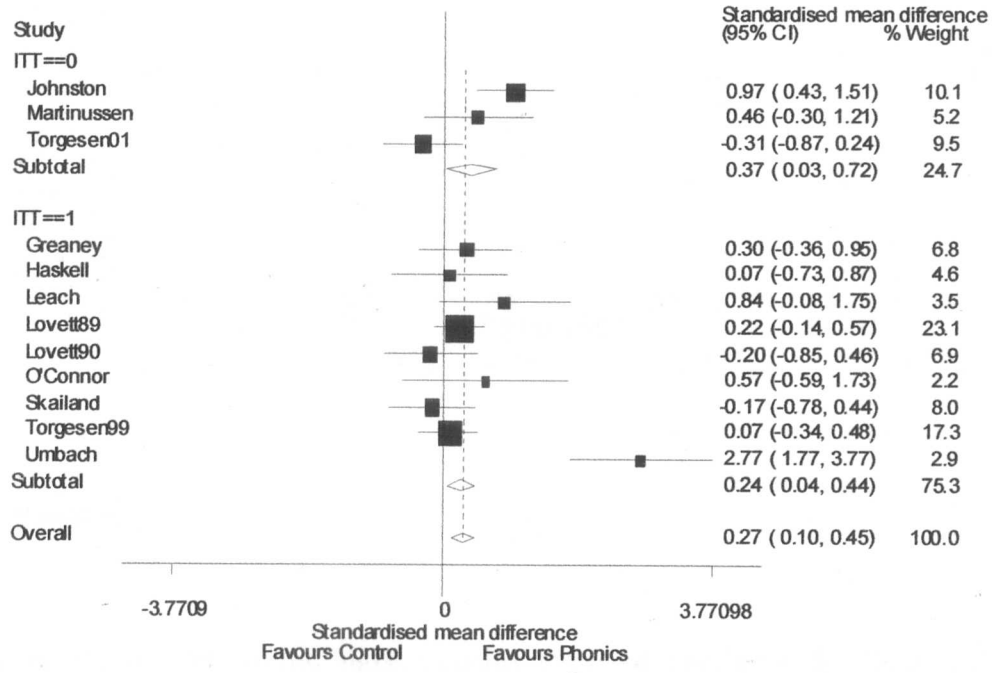


Figure 4.4 shows the meta-analysis subdivided by whether the children were normally attaining or had reading disabilities or difficulties (ability 0 refers to studies where the participants had leaning difficulties or disabilities; ability 1 refers to studies where the participants were normally attaining). As the forest plot shows, phonics instruction for normally attaining children tended to produce a greater effect size of 0.45 compared with children with reading disabilities and difficulties (0.21). However, the test for interaction was not statistically significant (p=0.24). Therefore, there is no statistical evidence to support the belief that the effectiveness of phonics instruction is materially different between learners with different characteristics.

There are other potential, methodological, sources of heterogeneity. One such source could be whether or not studies used intention to teach analysis. In Figure 4.5 an analysis is shown of whether or not studies differed in their results by the use of intention to teach (ITT 0 refers to studies where ITT was not used; ITT 1 refers to studies where ITT was not used). As the figure shows, studies that used ITT analysis

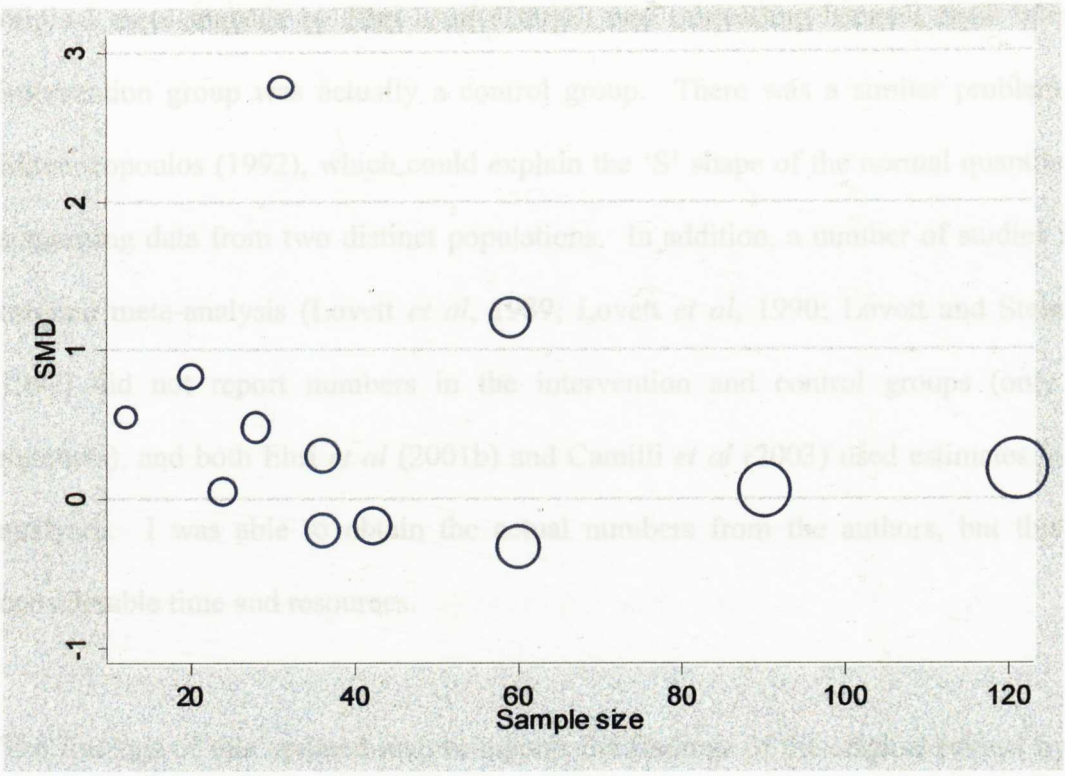
tended to have smaller effect sizes; however, this apparent interaction was not statistically significant (p=0.72). Tests for interaction tend to have low power and, given the relatively small number of studies in this review, we cannot conclude that failure to use ITT did not affect the effect size.

Figure 4.5: Meta-analysis subdivided by ITT or no-ITT



Finally, I looked at whether there was any relationship between the sample size of the study and its effect size. In figure 4.6 I present the results of a 'bubble' plot which plots the size of the study by its effect size. As the figure shows, there appears to be little or no correlation between the sample size of the study and the standardised mean difference.

Figure 4.6: 'Bubble' plot of effect size and sample size.



Discussion

In this chapter one of the highest quality systematic reviews identified in the tertiary review was updated. The original meta-analysis was suggestive both of publication bias, when inspected using a funnel plot, and the possibility of inappropriate pooling of studies indicated by the normal quantile plot (see the previous chapter). The aim of this chapter was to investigate the issue of inappropriate pooling and to see if, by extending the search strategy to include unpublished studies, the possible problem of publication bias could be reduced.

Replication of the meta-analysis did find inappropriately included studies. For example, Gittelman and Feingold compared a 'whole language' instruction which contained some phonics with a classroom survival strategy, whereas the majority of studies compared

phonics instruction versus other forms of literacy teaching. Including this study in the original meta-analysis by Ehri *et al* (2001b) was misleading because their 'phonics' intervention group was actually a control group. There was a similar problem with Mantzicopoulos (1992), which could explain the 'S' shape of the normal quantile plot, suggesting data from two distinct populations. In addition, a number of studies in the original meta-analysis (Lovett *et al*, 1989; Lovett *et al*, 1990; Lovett and Steinbach, 1997) did not report numbers in the intervention and control groups (only total numbers), and both Ehri *et al* (2001b) and Camilli *et al* (2003) used estimates in their analyses. I was able to obtain the actual numbers from the authors, but this took considerable time and resources.

The findings of this updated review support the findings of the original review by Ehri *et al* (2001b). However, there are some important differences. The overall effect size of 0.27 was substantially reduced compared to the Ehri *et al* (2001b) estimate of 0.41, which makes the use of phonics instruction seem less beneficial than originally supposed. The reduction in the effect size in the updated review can be explained by the inclusion of new trials from the updated searches and the original review's inappropriate inclusion of some trials, its use of what was essentially an untaught control group as the counterfactual in some comparisons (which is likely to exaggerate the effects of phonics teaching), and the lack of adjusting for the clustering effects in the calculation of the mean effect size in the one cluster trial (Brown and Felton, 1990).

There is also a significant amount of heterogeneity in the review, which is not explained by any obvious source in methodological or educational factors. Failure to find reasons for heterogeneity is an issue for concern as this may be due to poor methods exaggerating or underestimating the effectiveness of phonics instruction or the method

of delivery of phonics or the population to whom it is delivered. The lack of detailed description of the trial methods in most studies precludes exploration of methodological issues.

Quality issues

None of the 14 trials in the updated review reported method of random allocation, sample size justification or blinded assessment of outcome. Nine of the 14 trials used intention to teach (ITT) analysis. Whilst this seems to be good, it could be explained by the fact that some educational researchers do not routinely report attrition and imply that there were no drop-outs (which may not in fact be the case).

Replicability of the original review is key for judging Ehri *et al*'s (2001b) findings, but two further independent reviewers or teams of reviewers came to different conclusions. According to quality appraisal of this review using the QUORUM statement (see Item 2) it was a comprehensive and rigorously designed systematic review of the effectiveness of systematic phonics instruction, yet there were a number of problems associated with its replication, e.g., inappropriately including some trials, the lack of detailed quality appraisal of included studies, and not stating that numbers had to be estimated.

Conclusions

This reappraisal of the NRP review has found that systematic phonics instruction is associated with an increased improvement in reading accuracy. The effect size is 0.27, which translates into an approximate 12% absolute improvement in a reading accuracy test that is standardised to have a score with a mean of 50% for children not receiving

systematic phonics (see Torgerson, 2003, Item 1, p 86). In other words, for 100 children not receiving systematic phonics instruction 50 would score 50% or more, compared with 62 children who would get 50% or more if they received systematic phonics instruction. Such a difference is probably educationally worthwhile. Nevertheless, the findings do need to be treated with caution. There is significant heterogeneity within the meta-analysis, which could not be explained by the *observable* (or reported) design characteristics of the included trials or by the educational characteristics of the children included in the studies. Therefore, it is unclear whether systematic phonics teaching is beneficial to *all* children with different learner characteristics. Only one of the included trials was undertaken in a UK context, which raises concerns about the applicability of the results to a UK setting. In addition, there is the issue of publication bias. The strong possibility of publication bias affecting the results cannot be excluded.

Two conclusions are drawn from this chapter. The first is that, for systematic reviews that have major educational implications, a replication of the review by a second independent research team is warranted. The second is whether or not exclusive, widespread use of systematic phonics in early literacy development as recommended in the UK by many stakeholders (policy makers, politicians, teachers) should be implemented. My conclusion is that a more cautious approach is justified. This approach would require a research team to design and undertake a large pragmatic RCT within a UK setting. Such a trial would probably be of a cluster design with schools being allocated to either maintain current adherence to the Primary National Strategy (the control group), or to receive additional systematic phonics instruction or replacement of the PNS by exclusive systematic phonics teaching (the intervention groups). Such a trial, that would have sufficient power to demonstrate a difference of

0.27, would probably need between 150 to 200 schools, i.e., several thousand children. Whilst such a trial would be expensive, it would be relatively cheap compared with the cost of inappropriately implementing early and exclusive systematic phonics instruction to all children in the UK in the Reception years, i.e., getting the national literacy policy wrong.

Section 3

Design bias

Chapter 5: Design bias in randomized controlled trials

Introduction

The randomized controlled trial (RCT) is the most robust study design for assessing effectiveness. Randomization eliminates selection bias and ensures that a study has high internal validity, i.e., a high degree of causality can be inferred (Cook and Campbell, 1979; Shadish *et al*, 2002) and so lack of randomization is associated with bias (Begg and Berlin, 1988). This is because groups formed by randomization will, on average, have similar characteristics, and therefore alternative explanations for the results can be ruled out. Selection bias may occur in non-randomized controlled studies because those participants receiving the intervention may be intrinsically different from those participants in the control group and this can give potentially misleading results.

Systematic review methods are particularly valuable when the field of inquiry contains large numbers of relatively small, randomized trials. When examined on an individual basis, small, randomized trials can give misleading results. This is because small studies are less precise: they have relatively low statistical power to detect modest but important differences in educationally significant outcomes. By using meta-analytical methods similar trials can be pooled to enable the analyst to observe, as statistically significant, worthwhile effects that individual trials may have missed.

Whilst meta-analysis can go some way towards addressing the problem of underpowered trials, it will not produce a true estimate of effectiveness if the trials contained within the analysis are methodologically flawed. Non-randomized studies,

studies of poor quality (for example RCTs that fail to conceal allocation) and studies with small sample sizes may result in exaggerated effect sizes (Lipsey *et al*, 1985).

Wilson and Lipsey (2001) examined a database of 319 meta-analyses of psychological, behavioural and educational treatment research to investigate the extent of the variance associated with study design. They found that the randomized designs yielded slightly higher effect sizes, that higher quality studies tended to have smaller effect, sizes and that larger sample sizes tended to yield smaller effects.

Quality assessment of studies included in a meta-analysis

Glass (1977) originally advocated the inclusion of all the available studies in a field in a meta-analysis regardless of differential methodological quality of the studies. This was criticised, and Glass *et al* (1981) subsequently stated that all the studies should be included but that they should all be quality appraised. This was in order that any interaction effect between study quality and effect size could be investigated. Slavin (1984) subsequently claimed that investigation of this relationship was rarely evident in meta-analysis and suggested that studies should be selected for inclusion in meta-analyses based on an assessment of their methodological quality.

The CONSORT statement

Over the past decade methodologists working in the health field have developed a set of guidelines that trialists should adhere to in order to ensure high quality reporting of RCTs. These have been published as the CONSORT statement (Begg *et al*, 1996; Altman *et al*, 2001). The motivation for CONSORT was the poor quality of so many of the RCTs that have been published in the healthcare field. Methodological reviews have highlighted the weaknesses of many randomized controlled trials of healthcare

interventions published in major medical journals (Altman and Dore, 1990; Moher *et al*, 1994; Schulz *et al*, 1995). Healthcare and educational trials face similar methodological challenges and are likely to suffer from similar methodological flaws.

Part of the need to develop the CONSORT statement arose from the widespread use of systematic review and meta-analysis methods in healthcare research. Ideally, in a meta-analysis, trials using similar methods need to be identified for the purposes of 'pooling' the results. In order to identify a sample of homogeneous trials the trial methods need to be clearly and correctly described. As well as allowing the identification of similar trials, CONSORT allows meta-analysts to exclude studies that are so methodologically weak as to be potentially fatally flawed.

Whilst, arguably, trials are undertaken in healthcare research more often than in any other field, the use of controlled trials in other areas has a longer history. Indeed, agricultural trials pre-date healthcare trials by many years (Oakley, 2000). Similarly, in educational research trials have been widely used for a long period of time (Oakley, 2000). The first book to clearly describe the statistical and methodological approach to the design and analysis of randomized controlled trials was by Lindquist (1940) in the context of educational research. This book preceded the first reports of healthcare trials by several years. It is of interest that, in the preface to that book, Lindquist argued that a motivation for the book was to introduce trial terminology to educational researchers unencumbered by the jargon of 'agricultural statistics'. Healthcare research was not mentioned.

The quality of educational trials is likely to become of increasing importance if, and when, educational policy-makers start to demand evidence for policy changes based on

randomized data. Indeed, there are increasing calls from some educational researchers to undertake high quality randomized trials (see, for example, the call in Hannon, 2000 for the UK Government-funded evaluation of the pilot scheme of the National Literacy Strategy at Key Stage 3 to be tested using a randomized controlled trial; see also Boruch, 1994; Oakley, 2000; Torgerson and Torgerson, 2001; Torgerson *et al*, 2003). When trials are pooled in a meta-analysis it is important that the characteristics and methods of the individual trials are clearly described (Torgerson, 2003).

Although the RCT should, in theory, eliminate selection bias, there are instances where bias can and does occur. If there is bias within an RCT this will invalidate the design and make the results no more reliable than an observational study. Indeed, a biased RCT may be more damaging to evidenced-based policy-making than an observational study, as the latter is acknowledged as having threats to its internal validity. Subsequent interpretation is guided by this knowledge.

Randomization minimises or controls for biases when groups are selected into a trial (Chalmers, 2001). If control groups are formed inappropriately this can introduce systematic bias (Begg and Berlin, 1988). Some educational researchers have recognised that randomization *per se* does not guard against all forms of bias (for example, randomization does not by itself control for bias that can occur when outcomes are assessed: observer bias) and have sought methodologies to improve the internal validity of their trial design.

Sources of bias in randomized controlled trials

The aim of this chapter is to discuss some of the sources of bias that can occur in randomized trials in educational research, and to outline how these biases can be minimised. The biases that Chapter 7 explores are dilution bias; detection or reporting bias; design bias; attrition bias; subversion bias; and exclusion bias.

Dilution or performance bias

Dilution bias can occur because individual trials may allow 'leakage' of the intervention to the control group, possibly when participants or their teachers or parents differentially seek alternative treatments after randomization. For example, in the case of a trial evaluating systematic phonics instruction to improve reading scores, it is possible that the parents of children not allocated to the phonics intervention might deliberately seek out the instruction elsewhere. This will have the effect of producing a biased under-estimate of treatment effect and increasing the possibility of a Type II error: that is, erroneously concluding there is no difference, when in fact there is a difference. This problem can be reduced by, if possible, blinding the teacher, parent and participant to the intervention. In addition, or alternatively, the control intervention can be made relatively more attractive than the intervention, and this may reduce the problem.

Alternatively, dilution bias can occur if the intervention treatment is inadequately delivered to all the children in the active group. This will dilute the observed effect of the intervention.

Trials ideally should report the methods for reducing the possible threat from this bias. One approach to dealing with the problem of dilution effects is to use cluster or group randomization. If pupils are randomized by class to different interventions this reduces the risk of contamination to the control group. An example that is possibly unique in education empirically tested the effects of leakage or contamination via a randomized trial that allocated both at the cluster level and then, within the intervention group, also allocated individual pupils. The intervention students received additional teacher praise in mathematics and reading, while the control children received normal feedback. The outcome measure was score on an academic self-concept scale. The intervention resulted in higher scores on the self-concept scale for internal controls than external controls, providing evidence of leakage of the intervention between control and intervention pupils among the individually randomized pupils (Craven, 2001).

Detection or reporting bias

Detection, reporting or assessment bias occurs when researchers are more assiduous in their reporting of events in one group compared with the other group. Reporting bias can be minimised if reporting is undertaken 'blind' to treatment allocation (Cook and Campbell, 1979), although this is not always possible. Alternatively, event reporting can be ascertained from differential sources (e.g., participant and teacher and local authority), which should minimise the risk of detection bias. Therefore, a good quality trial should report that outcomes were either assessed blindly or were assessed in some independent manner.

Design bias

The choice of a sample size of any trial is very often an arbitrary process. One of the key issues in the use of a sample size calculation is the potential to minimise the

chances of making a 'Type II' error. A Type II error occurs when there is a true difference between the randomized groups; however, this difference is either not statistically significant or is simply missed. The probability of a Type II error occurring declines with increasing sample size: larger trials, other things being equal, are less likely to experience a Type II error than smaller studies.

Typically, we want to have a sample size that will have a low probability of missing an educationally important difference. One of the key problems with respect to sample size estimation is the definition of what difference is important. For the purposes of this chapter it is proposed that a half a standard deviation difference (i.e., 0.5 effect size) is the largest difference that we should power our trials to detect. The justification for this is that an overview of quasi-experimental research in the educational and psychological literature noted that most interventions had an effect size of 0.5 or smaller (Lipsey and Wilson, 1993). Recent systematic reviews I have undertaken in the fields of volunteering, spelling and ICT and literacy confirm that few trials show an effect that exceeds 0.5 of a standard deviation (Torgerson *et al*, 2002; Torgerson and Elbourne, 2002; Torgerson and Zhu, 2003). Few commonly evaluated interventions give larger educational effects. Similarly, if the outcome is a binary variable it is unlikely that any intervention will lead to greater than a doubling or halving of effect.

The other issues that affect sample size are: power and significance. Traditionally power is often set at 80%, with 5% significance. Therefore, as a minimum a RCT should be powered to detect half a standardised difference between two groups, or a halving or doubling of a dichotomous outcome. For continuous outcomes, using individual randomization, detection of an effect size of half a standard deviation between two groups at this power requires 126 participants. Therefore, ideally,

educational trials with two groups should be designed to have at least this number of participants.

Sample size also can act as a proxy for overall trial quality. Researchers who recognise the importance of having large samples within their trials are usually the same methodologists who recognise the importance of limiting bias within their trial. In a methodological review of healthcare trials, Kjaergard and colleagues (2001) noted that small trials tended to be of poorer quality than larger trials. It is likely that the same will apply to educational trials.

Attrition bias

Most trials lose participants. Unless attrition is a random event, any amount of attrition can lead to the possibility of selection bias because participants with different characteristics may leave one arm of the study preferentially. This possibility is minimised, but never eliminated, if attrition rates are similar between the arms of the trial. Attrition bias is minimised if assiduous follow-up methods are applied to all randomized participants. A robust report of a randomized trial should always describe the attrition rate of its participants by allocated group, even if this is simply to state that there has been no attrition. Trials with attrition should undertake a sensitivity analysis of their results to see if these change when lost participants are included. This can be done using a best- or worst-case scenario. For example, if an intervention is effective, does it remain effective if those who are lost in the intervention group are assigned the worst scores and those in the control group are assigned the best scores? If the results remain positive and statistically significant then we can be confident that the results are reliable.

Subversion bias

Subversion bias is when researchers deliberately or unconsciously preferentially allocate participants with certain characteristics to one of the treatment arms. Clearly, if participants are allocated into their groups other than in a random fashion, bias will occur. Anecdotally, in healthcare, this practice has been relatively widespread (Schulz *et al*, 1995; Schulz and Grimes, 2002). Moreover, the effects of subversion bias have been demonstrated statistically. When the effect sizes of trials that contain measures to prevent subversion are compared with studies that do not have such measures, the effect sizes of the former are lower (Schulz *et al*, 1995). For instance, in trials where the allocation sequence is made known to researchers in advance of randomization, the effect sizes will tend to be exaggerated compared with trials where the allocation sequence is concealed from the researcher until the moment the participant is randomized (Schulz *et al*, 1995; Chalmers, 2001). Boruch (1997) reported an instance of subversion within a criminal justice trial, with police officers apparently over-riding random assignment in an evaluation of a domestic violence prevention programme. Whilst I am unaware of any reports of subversion of allocation schedules in the educational literature on trials, it is highly unlikely that educational trials are immune from the problem. Therefore, randomization needs to be separated from the researcher who is recruiting the participants. This can be achieved by using an independent researcher to randomize the trial participants and/or by using a computer to generate the randomization schedule. A hallmark of a good report of a trial is a description of independent allocation, which is always possible and indicates that scrupulous compliance with the design of the study.

Exclusion bias

It is a feature of many studies to exclude participants after randomization. This occurs for a number of reasons apart from attrition. For instance, some participants may not have received the intervention. Including such participants in the main analysis will result in dilution bias: thus, the effect of treatment will be underestimated. Nevertheless, to exclude these participants will introduce another bias, because those excluded are unlikely to be representative of the sample as a whole and their counterparts in the control group will be retained in the analysis. Ideally, once participants have been randomized they should be retained within the analysis to avoid this bias: that is by intention to teach.

Discussion

This chapter has summarised some of the main biases that can affect the validity of RCTs. Many, if not all, of these biases can be avoided through careful attention to the design and execution of the RCT. Note it is important to emphasize, whilst all these biases can and do occur in RCTs, they will also occur in other study designs (e.g., pre- and post-test methods).

The next chapter, Chapter 6, undertakes a review assessing the prevalence of these biases firstly within a general sample of educational trials, and then within a specific sample of literacy trials.

Chapter 6: Assessing the quality of randomized controlled trials in educational research

Introduction

Researchers undertaking systematic reviews in healthcare have developed several quality appraisal checklists (e.g., the Jadad scale, Jadad *et al*, 1996), most notably the CONSORT statement (Begg *et al*, 1996; Altman, 2001). Measurement of study quality is not necessarily an objective exercise. The use of any quality score can be fraught with difficulty. For example, Juni and colleagues quality appraised 17 healthcare trials with 25 different quality scales (Juni *et al*, 1999). They found that for 12 scales the effect sizes were the same when trials were rated as high or low quality. However, for six scales, high quality trials showed little or no benefit of treatment compared with low quality studies, whilst the remaining seven scales showed the opposite result. Thus, quality assurance scales can give very different results depending on the items included and the weights given to individual items. Even describing a study as being randomized or not can be difficult to ascertain in some evaluations. For example, a phonological awareness study by Hatcher and colleagues (1994) does not appear to be a randomized controlled trial in the published paper. Therefore, if one relied solely on the published paper, such a study would be assessed as a controlled study, not the more rigorous randomized trial. However, the children in the study were actually allocated to their treatment groups in a randomized fashion (Hatcher, personal communication, 2001). Similarly a widely cited controlled trial comparing different approaches for the teaching of phonics (Johnston and Watson, 2004) does not state in any of its published reports how the children were allocated to their teaching groups. Only after correspondence with the lead author was it established that the allocation procedure for the first of the

two experiments was researcher decision, while in the second experiment matching was followed by random allocation (Johnston, personal communication, 2005); thus the two trials were respectively non-randomized and randomized.

Assessment of trial quality is important in order to allow the reviewer to make a judgement about whether or not each trial is of sufficiently high quality to be included, and to enable the readers of the systematic review to judge the quality of any included trial.

The CONSORT statement

The CONSORT statement is a consensus statement originally developed by leading health trial methodologists in 1996 (Begg, 1996) to help to improve the quality of reporting of trials in healthcare research. In 2001 the CONSORT statement was revised (Altman, 2001). It is not a quantitative scale. Rather it is a checklist and flow diagram for trialists on how to report their trials to enable reviewers to ascertain whether the most important aspects of a trial design have been fulfilled. Lack of a scoring mechanism may be seen as an advantage of the CONSORT approach as it allows the reviewer more flexibility in the interpretation of a trial's quality. The revised CONSORT statement is a 22-item checklist, including items examining the quality of reporting of the rationale, design, conduct, analysis and interpretation of RCTs.

Quality assessment criteria in educational research

Quality assessment criteria have been used previously in educational research and were identified in the process of undertaking the searches for reviews to include in the tertiary review (Item 2). The tertiary review identified three systematic reviews that undertook methodological analyses of experimental research in two literacy areas: comprehension

strategy instruction (Lysynchuk *et al*, 1989; Ridgeway *et al*, 1993) and phonological awareness training (Troia, 1999). However, these reviews were not included in the tertiary review because they were methodological reviews, and did not contain pooled effect sizes. The two reviews in comprehension strategy instruction are briefly described, and the third review in phonological awareness training provides the justification for, and the basis of the development of, the criteria that are used in the methodological analyses in Chapter 7.

Comprehension strategy instruction (Lysynchuk et al, 1989)

Lysynchuk *et al* (1989) evaluated the reading comprehension instructional studies according to a broad range of criteria based on generally accepted standards of internal and external validity (Campbell and Stanley, 1966). Thirty-seven studies were critically evaluated in the paper using 24 internal validity criteria and 6 external validity criteria. The authors concluded that, in terms of internal validity criteria, although many of the studies had important strengths, many also had serious weaknesses. The authors identified a number of confounding variables in the 37 studies which could have been avoided by better-designed studies. The confounders included: lack of random assignment to experimental or control conditions; Hawthorne effects (Campbell and Stanley, 1966); lack of control group exposure to the same training materials as the experimental group; lack of information about the amount of time experimental and control subjects spent on dependent variable tasks; and use of inappropriate unit of analysis. In terms of external validity criteria, the authors concluded that the majority of the studies met these criteria. However, they also noted that most of the studies did not assess long-term effects or transfer of the strategy instruction. Finally, the Lysynchuk *et al* (1989) noted that some of the most seriously methodologically flawed studies have

subsequently been very influential in both theory and practice, for example, including reading strategy instruction 'as a prominent component' in basal reading series.

Comprehension strategy instruction (Ridgeway et al, 1993)

Ridgeway *et al* (1993) re-analysed 98 studies included in a review conducted by Alvermann and Moore (1991) on secondary reading practices (teaching and learning strategies), in order to examine the methodological quality of the policy and practice research base. They used the internal and external validity criteria developed by Lysynchuk *et al* (1989), in turn developed from Cook and Campbell (1979). Again, the authors concluded that the research base included instances of methodological strengths and weaknesses. The main weaknesses highlighted occurred in the data collection processes (choice of materials, training of subjects and measurement issues); and in the data assumptions in the statistical analyses.

Phonological awareness training (Troia, 1999)

Troia examined the methodological quality of 39 studies of phonological awareness interventions. In this review the author developed a set of quality criteria in order to judge the quality of the included randomized and controlled trials. These criteria were later adopted by another review that was included in Item 2 of the portfolio, the tertiary review (Ehri *et al*, 2001a). The internal validity criteria included design characteristics and statistical analysis; the external validity criteria included generalization issues. Again the most serious design flaw was deemed to be non-random assignment to experimental or control groups. Other confounders highlighted included failure to control for Hawthorne effects and inadequate sample sizes.

Two of these methodological reviews concluded that the most serious threat to the validity of the reviews was lack of random assignment, following Cook and Campbell (1979) (Lysynchuk *et al*, 1989, p. 462; Troia, 1999, p.48). The reason for this is that random assignment is the only design feature that can control for all selection biases, except chance bias (Cook and Campbell, 1979; Shadish *et al*, 2002). Therefore the methodological work undertaken in the remainder of this chapter focuses solely on important design features of randomized controlled trials.

Methods

It was decided to critically examine one of the three quality appraisal sets of criteria that had been developed in literacy research. Troia's was selected because it was the most recent. Rather than adopting Troia's tool uncritically for quality assessment of educational trials in general, it was deemed to be important to compare the characteristics of this tool with another tool in order to test the validity of the instrument.

One of the aims of this chapter, therefore, is to take elements of good practice and apply them to a selection of RCTs: the trials that both Troia and I agree are randomized controlled trials.

Description of quality tools

The two quality criteria to be compared are those used by Troia (1999) and an amended version of the CONSORT statement. Both will be applied to a sample of RCTs. The sample of RCTs is taken from Troia's (1999) systematic review.

The Troia quality criteria

Troia developed his quality assurance criteria by scoring each dimension of quality on a three-point scale, with higher scores indicating lower quality. In Table 6.1 I summarise Troia's quality criteria and comment on each criterion. Apart from Troia's first and last validity criteria (random allocation and equivalent mortality rates) all other criteria are not essential in order for a trial to be internally valid. In addition Troia gave arbitrary weights to his criteria. One problem with giving weights to criteria is that a study can have some fatal flaws in terms of its internal validity but yet score quite highly because it fulfils all the other criteria, thus giving a potentially misleading impression of trial quality.

Table 6.1: Troia's study quality criteria relating to internal validity

Troia's validity criterion	Troia's weighting	My comment(s)
Random Assignment	3	Most important internal validity criterion.
Control group received alternative intervention to control for Hawthorne effect	3	Not necessary to achieve internal validity if the comparison is against standard practice.
Exposure to similar materials for control group	1	As above, not necessary for internal validity.
Counterbalancing of teachers	2	This is necessary, although a single instructor could introduce bias if he/she had conscious or unconscious preference for one treatment. To avoid this, multiple instructors teaching each condition are required.
Treatment explicitly described	2	Not an internal validity requirement; is a requirement for replication and implementation of intervention.
Criterion based intervention	1	Not necessary for internal validity.
Equivalent instructional time	3	Depends on the study question; not necessary if comparing an intervention that requires a different length of instruction against 'standard practice'.
Equivalent mortality rates	1	Very important for internal validity; differential mortality or attrition rates can lead to bias.

Some of Troia's choices of quality criteria and the weights given are inconsistent with the views of other trial methodologists (e.g. Begg *et al*, 1996). For example, Troia allocated a weighting of 3 (fatal flaw) if the control group had not received the same instructional time. If the study is a pragmatic trial (Schwartz and Lellouch, 1967;

Torgerson and Torgerson, 2001) then requiring the control group to have exactly the same instructional time may be unnecessary. The design of a pragmatic trial mimics 'real life' classroom conditions as closely as possible. Therefore, a novel intervention may be longer or shorter than the standard practice it seeks to replace. Attempting to 'artificially' increase or shorten the length of standard teaching to ensure equivalent length of teaching ensures that the results of the novel teaching method are only superior or inferior when compared with a teaching practice that is not standard practice. In other words the trial is evaluating at least one 'artificially' contrived treatment condition, which would not be used in a typical teaching situation. Results from such a trial, therefore, may not be very relevant to practical teaching. It is far better to adopt standard practice as the control and accept this even if this is longer or shorter than the intervention.

Troia also identified some cluster trials, where pupils had been randomized in groups (e.g., by class); however, he categorised these as quasi-experimental studies, not as randomized trials. In principle, a cluster trial is as rigorous as an individually randomized study and should be classed as such. As long as there are sufficient numbers of clusters within a cluster trial to enable any differences between clusters to be balanced out through randomization, then randomizing by cluster will produce equivalent groups (Ukoumunne *et al*, 1998).

Attrition rates

Attrition, that is loss of children before follow-up, is a fact of life in many trials. Attrition is a problem in two respects. It can lead to selection bias and threaten the internal validity of the trial. It also reduces the power of the study. Troia recognised

the importance of attrition, but only penalised studies with differential attrition by a weighting of 1.

Intention to teach analysis

Once a fair randomization has occurred it is still possible for there to be bias within the trial if all those who are randomized are not included in the analysis of the groups to which they were originally assigned. A common but inappropriate analytical technique is the process of 'on treatment analysis' where children who do not receive the full intervention package are excluded from the study. Another popular inappropriate analytical strategy is to exclude children from the analysis if their pre- or post-test values are perceived to be 'extreme'. For example, authors of one of the studies identified by Troia excluded a pupil from their analysis because the pupil's change score was too great (Hatcher *et al*, 1994). Troia only partly addressed the issue of intention to teach analysis by penalising non-equivalent mortality rates by giving it the lowest weighting of 1.

Active treatment analysis or the exclusion of outliers can be justified as a post hoc exploratory or sensitivity analysis. If the results of the study do not fundamentally change using these other analytical techniques, then one can be more confident of the results. However, if there is a change in the results, particularly in the direction of effect, then this is a marker for the need to replicate the study.

Allocation concealment

Allocation concealment prevents 'subversion bias'. The extent of this bias has been well documented in the area of healthcare research and Boruch has noted an occurrence in criminal justice (1997). Troia, however, did not consider this issue at all in his

quality criteria. It is likely that educational trials are as susceptible to subversion effects as health and criminal justice studies. It is, therefore, essential that random allocation is concealed and undertaken by an independent person.

Blinded assessment of outcome

Ascertainment bias occurs when the researcher who collates the outcomes of an experiment either consciously or subconsciously selects better outcomes for students in the experimental group compared with the control arm. This has long been recognised as a serious bias in research (Cook and Campbell, 1979; Chalmers, 2001). A way of avoiding this form of bias is to have 'blinded' assessment of outcome: the person administering and marking the final post-test is composition of the intervention and control groups. If the researcher is not blinded s/he may inadvertently preferentially give higher scores to children who have been allocated to the novel intervention. Troia did not include blinded assessment in his quality criteria.

Non-internal validity criteria

Whilst internal criteria are the most important aspects of trial design, as described in Table 7.1, other important aspects need to be considered when designing a study.

Sample size

In the field of education most experimental innovations yield small to moderate positive effects (Kulik and Kulik, 1989). Therefore, researchers seeking statistical significance must use large sample sizes. Small or underpowered trials are wasteful of research resources. Moreover, they can lead to the wrong conclusions.

Troia, whilst recognising the importance of minimising the probability of a Type II error, weighted it relatively minimally (i.e., 1 point). He argued for a minimum sample size of only 10 per group, which would have a very low probability of being able to detect a difference of half a standard deviation between the two groups. Indeed, such a sample size of 20 in a two-group comparison would have a low probability of detecting 1 standard deviation between two groups, which requires a group size of at least 16 per group (i.e., 32 in a comparison between two groups).

Confidence intervals

No trial of phonological awareness instruction identified by Troia reported confidence intervals in its results section. The point estimate of effect from any trial is bounded by uncertainty. For instance, a large effect can be statistically non-significant because the sample size is too small. One way of representing the boundaries of uncertainty around the point estimate of effect is to use confidence intervals (usually 95%).

In summary, Troia either did not consider, or allocated low weightings to, three important factors, each of which can be a fatal flaw and result in a trial producing a biased answer. These were: allocation concealment; blinded assessment of outcome; and intention to teach analysis.

Modified CONSORT criteria

The methodological strengths or weaknesses of the phonological awareness instruction trials were re-assessed using some additional criteria from the CONSORT statement. These included concealed allocation and blinded outcome assessment, both of which were missing from the original Troia criteria.

Concealed Allocation

The unpredictability of random assignment is one of its key features and should not be compromised by allowing the researcher who is undertaking the randomization procedure foreknowledge of the next allocation (Schulz and Grimes, 2002). Cook and Campbell (1979) have observed that 'random allocation' is less likely to be random when people who do not understand the importance of random allocation undertake the randomization.

Random allocation *must* be made as unpredictable as possible for the researcher who is allocating the trial participant. Therefore, one hallmark of a good trial is attempting as far as possible to conceal the random allocation. Whilst concealed allocation is very important, Troia did not consider this issue at all in his quality assessment.

Results: re-assessment of the phonological awareness trials

In Table 6.2, results of this re-assessment of the RCTs identified by Troia are presented. It was impossible to tell from the reports of one of these trials (Wise and Olson, 1995) whether or not they did in fact use random allocation to experimental and control groups; therefore this paper was excluded from the analysis. In addition, one of the studies was not in fact a trial (O'Connor *et al*, 1995). The intervention children were 'low skilled' and the control children were 'high skilled', therefore, this paper was also excluded from the analysis. Troia also identified two papers that had reported randomization by cluster. Kozminsky and Kozminsky (1995) claimed that their trial was an RCT but it was in fact an observational study as there were only two clusters. Troia quite properly labelled this study as a non-randomized trial. In contrast, however, Lie (1991) reported a 3-arm cluster trial where schools were randomized to the

intervention. Troia claimed, inaccurately, that this was not a RCT. However, in Table 6.2, the assessment is confined to all the studies both Troia and I agree are RCTs.

Table 6.2: Characteristics of randomized trials of phonological awareness instruction

Author (date)	Reporting of method of allocation (CONSORT 8 and 9)	Sample size justification (CONSORT 7)	Intention to teach analysis	Required sample size	Actual Sample Size	Blinded outcome assessment	Confidence intervals (CONSORT 17)
Ball and Blachman (1991)	NS	NS	NS	126	60	NS	NS
Bentin and Leshem (1993)	NS	NS	NS	126	41	Yes	NS
Byrne and Fielding-Barnsley (1991)	NS	NS	NS	126	128	NS	NS
Byrne and Fielding-Barnsley (1993)	NS	NS	NS	126	119	NS	NS
Byrne and Fielding-Barnsley (1995)	NS	NS	NS	126	120	NS	NS
Castle *et al* (1994), Experiment 1	NS	NS	Yes	126	30	NS	NS
Cunningham (1990)	NS	NS	NS	126	28	NS	NS
Foster *et al* (1994), Experiment 1	NS	NS	Yes	126	27	Yes	NS
Foster *et al* (1994), Experiment 2	NS	NS	NS	126	69	Yes	NS
Hatcher *et al* (1994)	NS	NS	NS	126	64	Yes	NS
Hohn and Ehri (1983)	NS	NS	Yes	126	16	NS	NS
O'Connor *et al* (1993)	NS	NS	NS	126	21	NS	NS

Slocum *et al* (1993)	NS	NS	NS	126	24	NS	NS
Torgesen and Davies (1996)	NS	NS	Yes	126	100	NS	NS
Torneus (1984), Experiment 2	NS	NS	Yes	126	38	NS	NS
Uhry and Shepherd (1993)	NS	NS	Yes	126	22	NS	NS
Vellutino and Scanlon (1987)	NS	NS	NS	126	30	NS	NS
Weiner (1994)	NS	NS	NS	126	36	NS	NS

NS = not stated

Table 6.2 presents an analysis of the remaining 18 individually randomized trials. It focuses on a number of serious threats to internal and external validity. It also assesses whether or not for each trial there was a high risk of a Type II error, i.e., having insufficient numbers to detect an educationally important difference.

As Table 6.2 shows, none of the 18 studies reported the method of random allocation and only four appeared to use intention to teach analysis. Both of these shortcomings, as described previously, are serious threats to internal validity.

In terms of avoiding a Type II error, again all of the studies had flaws. No study reported the underlying justification for choosing the sample size and only one trial had sufficient power to measure at least half of one standard deviation (126) (Byrne and Fielding-Barnsley, 1991). This trial had a sample size of 126. Due to attrition rates, the two follow-up studies were slightly underpowered (Byrne and Fielding-Barnsley, 1993, 1995). In addition, five of the trials did not have a sufficiently large sample size to detect one standardised difference between the groups (i.e. for 80% at a 5% significance level two groups of 16 pupils need to be compared).

Whilst the Troia (1999) and the subsequent National Reading Panel review (Ehri *et al*, 2001a) present some evidence that phonological awareness training could have a positive effect on reading acquisition skills, all of the trials have some methodological flaws. Pooling or meta-analysing large numbers of poor quality trials does not provide any more robust evidence than only one or two weak trials. Indeed, large numbers of weak trials can give spurious estimates of confidence in the results if they all contain inherent biases, which point to the same flawed estimate of effect.

In the preceding section some of the key issues with respect to the quality of randomized trials in educational research have been highlighted. A re-appraisal of the quality criteria developed by an educational researcher has shown that in the field of phonological awareness instruction the criteria are insufficient to judge the quality of the RCTs for major methodological flaws. In the next section a methodological study evaluating the quality of a wider sample of educational trials published since 1990 is described.

A methodological evaluation of education trials[10]

In this methodological evaluation it was decided to apply a simple set of methodological criteria to a sample of educational trials. The analysis of trials evaluating the effectiveness of phonological awareness instruction suggests that in this field the quality of trials is weak. The extent to which this applies to a wider range of educational trials is unknown. To ascertain whether or not similar problems affect educational trials more widely, a quality assessment of a different sample of trials was required. The aims of this piece of methodological work, therefore, were to describe the characteristics of educational trials, and to assess whether or not the quality of the trials improved over time.

Identifying a sample of educational trials

Two approaches were used to identify educational RCTs. Studies identified through completed systematic reviews that I have undertaken recently were included (Torgerson and Elbourne, 2002; Torgerson *et al*, 2002; Torgerson *et al*, 2003; Torgerson and Zhu,

[10] This section of the item has been published as part of a combined paper: **Torgerson C.J.,** Torgerson D.J., Birks Y.F., Porthouse J. A Comparison of the Quality of Randomised Controlled Trials in Education and Health, *British Educational Research Journal,* 2005: in press. See Appendix B.

2003). To obtain a wider variety of trials, these were supplemented by trials identified through hand searches of key educational journals. Twenty references from the following hand searched journals were included: *The British Journal of Educational Psychology* (10 references); *Educational Research* (3 references); *Journal of Research in Science Teaching* (7 references). To prevent a single journal dominating the dataset, only one article from the same journal in any year was included.

After piloting data extraction from a convenience sample of trials, it was decided that the following five markers of trial analysis, methodology and reporting quality would be sought: concealed randomization; sample size justification; adequate sample size; 'blinded' follow up; use of confidence intervals.

Results

Eighty-four education trials published in 43 different journals were identified. Table 6.3 presents the characteristics of the identified trials. 15 (17.9%) of the 84 trials were cluster-randomized trials, and the remaining 69 were individually randomized trials. Six (7.1%) of the trials used a dichotomous outcome; the remaining 78 used a continuous outcome.

No trial described its method of randomization, or gave a rationale for the choice of sample size. Only one trial used confidence intervals to describe the uncertainty around the observed effect size. A minority of trials were either adequately powered or reported that they had used blinded follow-up.

Table 6.3: Prevalence of methodological characteristics in a sample of trials

Characteristics	Trials (n = 84)
Rationale for sample size	0
Allocation concealment	0
Blinded follow-up	12 (14.3%)
Use of confidence intervals	1 (1.2%)
Adequately powered	13 (15.5%)

To assess whether there was any association between the reporting quality of a trial and publication date, I correlated the mean number of reported items by year of publication. There appeared to be a decline in quality (-0.21, p=0.06) of trials published over time.

In table 6.4, the relationship between year of publication and the presence/absence of the quality criteria is explored. Trials show a significant decline in the numbers of trials reporting blinded follow-up and a tendency for trials to decline in statistical power. For two of the quality criteria (concealed allocation and sample size justification) it was not possible to look at changes over time as no study reported these. In addition, only one study reported confidence intervals (CIs).

Table 6.4: Increased odds of characteristic per year

Characteristics	Trials (n = 84)
Blinded follow-up	0.76 (0.61 to 0.94) p=0.013
Adequately powered	0.95 (0.80 to 1.13) p=0.56

NB An odds ratio of less than 1.0 indicates a decreased reporting of that quality characteristic.

It is interesting to note that, among these educational trials, a significant reduction in the use of blinded follow-up occurred with time. Blinding or masking assessors is important because this avoids reporting or ascertainment bias and has been a hallmark of good trial design for many years (Cook and Campbell, 1979).

Very few educational trials met the pre-specified sample size. Sample size is an important consideration when designing a study. Small trials can and will miss important effects. Whilst meta-analysis can and does go some way to addressing the problem of small sample sizes by pooling similar trials, this is an imperfect solution for a number of reasons. One important problem with small trials is that those which produce, by chance, a negative or null result, are less likely to be published than those which report a positive result (see Chapter 2 on publication bias). This in turn will bias the results of a meta-analysis by either overestimating a positive effect or, worse, erroneously concluding that an intervention has a positive effect when in truth it does not. Larger trials on the other hand, whatever their results, are more likely to be published. They are also more likely to give an estimate that is closest to the 'correct' answer than smaller studies. The finding that 85% of a sample of educational trials did not have adequate power to show a result as being statistically significant (albeit at the arbitrary 5% value) is a cause for concern.

It is important to take into consideration that, apart from sample size and the use of confidence intervals, the other markers of trial quality are dependent on the authors' reporting. It is possible that, in some trials, important aspects of trial quality such as blinded outcome assessment were undertaken but were not reported by the authors. This problem of under-reporting of trial quality could be addressed by editors of educational journals developing a similar checklist to the CONSORT statement to quality assure the publication of educational trials. This would not only help those who review trials in education but also act as an incentive for future educational trials to be more rigorously designed and executed.

Conclusions

In this chapter I have described the design aspects of rigorously designed trials. I have undertaken two distinct pieces of methodological work. First, I have critiqued the quality criteria developed by an educational researcher and later adopted by the alphabetics subgroup of the National Reading Panel (Ehri *et al*, 2001a) for their meta-analysis of studies evaluating phonemic awareness instruction. I have shown that these criteria are insufficient and that, when modified criteria based upon the CONSORT statement are applied to trials in phonological awareness training, these show that many of those trials are methodologically weak. I have then gone on to apply a subset of these CONSORT criteria to a sample of 84 trials sampled from the educational literature over a 12-year period. I have shown that these trials like those in the field of phonological awareness instruction are methodologically weak. I have also shown that there has been a decline in methodological rigour over the time period.

In the next chapter a detailed exploration of whether the design characteristics of the trials are associated with the effects of the intervention will be undertaken.

Chapter 7: The relationship between trial design and its outcome; and an exploratory meta-regression of the effects of characteristics of trials on effect size[11]

Introduction

In the previous two chapters the importance of trial design was discussed, and the variable quality of randomized controlled trials within the fields of educational research generally and literacy learning in particular was demonstrated. Trial design quality has been shown to be associated with outcomes in many methodological reviews in healthcare. The seminal work in the field was probably by Schulz and colleagues (1995). The issue, however, has been revisited on many occasions with Kjaergard (2000) noting that the relationship between effect size and sample size could be explained by the fact that trials with small sample sizes tend to be of lower quality than trials with large sample sizes. When differences in quality are taken into account, the relationship between effect size and sample size tends to disappear.

In the field of social science research, Wilson and Lipsey (2001) found that higher quality studies tended to have smaller effect sizes, and that larger sample sizes tended to yield smaller effects.

In the area of literacy research no similar methodological work has been undertaken in a systematic fashion. The aim of this chapter, therefore, is to explore the relationship between the quality of a sample of trials identified in the field of literacy learning and their reported outcomes.

[11] A slightly revised version of this chapter has been submitted to *Scientific Studies in Reading*

Methods

This methodological component of the item uses data from the sample of trials identified as part of the tertiary review (Item 2). In the tertiary review 14 systematic reviews undertaken in the field of children's literacy learning between 1983 and 2003 were identified that included at least one randomized controlled trial. In seven of these reviews it was possible to identify the individual randomized trials; therefore these seven reviews provide the source of the individual trials for the analysis reported in this chapter.

Key items of data quality were extracted and tabulated from the seven reviews. Because of the experience of reviewing trials outlined in the previous section, it was decided not to identify and extract some items associated with trial quality. One of the more notable quality items excluded was the method of random allocation. It was not possible to identify in any of the 84 general educational trials or the 18 trials in Troia's (1999) review (analysed in Chapter 6) any detail of whether or not the allocation was concealed. Therefore, it was assumed that the same would be true of the literacy trials. Indeed, whilst data were being extracted from the studies in literacy research, no description of concealed allocation was noted. Similarly, only one trial from these two earlier samples was identified which used confidence intervals, and therefore these were also not sought. It was decided that the methodological items of the trials that were likely to be identifiable and have some prevalence were: sample size, unit of allocation (i.e., individual or cluster), whether or not follow-up was 'blind', whether or not intention to teach (ITT) analysis had been used, and the attrition rate.

In addition, only one effect size for each trial was extracted. A judgement was made with regard to the most important educational outcome for each aspect of literacy that was to be extracted prior to examining the effect sizes in each study. For phonemic awareness this was reading (word recognition) followed by phonemic awareness; for reading this was comprehension followed by reading level or grade; for writing this was holistic score or quality of writing; and for spelling this was spelling test or inventory. If an aggregate effect size had been calculated this was used, and in the case that none of the preceding effect sizes was available for an individual trial, the first available effect size was used. Many of these aspects of coding had been already data extracted by the authors of the systematic reviews. However, in all cases the original trial report was obtained and any relevant data were extracted from that. Effect sizes were not re-calculated: in all cases an effect size (as described above) that had been calculated by the original reviewer was selected.

In this chapter I will examine the characteristics of trials that might affect the outcome of the study. The hypotheses to be tested in this chapter are:

(1) Cluster trials will produce effect sizes that are larger than individually randomized trials due to lack of contamination of the control group;

(2) Trials reporting blinded administration of follow-up measures will report smaller effect sizes than trials that do not report this (this is because blinding reduces the possibility of ascertainment bias);

(3) Trials with high drop-out, or attrition, rates will produce, on average, higher effect sizes due to bias;

(4) Trials that use intention to teach will have smaller effect sizes than those that do not;

(5) Unpublished trials will have a smaller effect size than published studies.

Statistical issues

To explore the relationship between trial characteristics and effect size I first undertook simple descriptive statistics comparing the mean effect size using weighted ordinary least squares regression. It is necessary to weight the different studies because each study has a different sample size and therefore those studies with the largest sample size should have the greatest weight in the analysis. To undertake a multivariate analysis I have extended this approach by including all the variables in the analysis. For the weighting factor I used the sample size of the study. Because cluster trials need to be treated differently with respect to sample size I estimated an intra-cluster correlation coefficient from a recent RCT of ICT and spelling (Brooks *et al*, 2005) which was 0.45. I then applied this to the formula: 1+(m-1) x ICC, where m is the average size of the cluster. This allowed me to calculate an effective sample size after adjusting for the clustering effects.

Results

Seven systematic reviews identified in the tertiary review allowed for the identification of individual randomized controlled trials (Bus and van IJzendoorn, 1999; Ehri *et al*, 2001b; Gersten and Baker, 2001; Mathes and Fuchs, 1994; Torgerson and Elbourne, 2002; Torgerson *et al*, 2002; Torgerson and Zhu, 2003) and were included for the purposes of the methodological work described in this chapter. A total of 56 randomized controlled trials were included in these seven systematic reviews. A number of trials appeared in more than one systematic review. After de-duplication the numbers of trials identified from each of the reviews was as follows (Table 7.1).

Table 7.1: Included trials

Author (date)	No. of RCTs identified
Bus and van IJzendoorn (1999)	14
Ehri *et al* (2001b)	12
Gersten and Baker (2001)	3
Mathes and Fuchs (1994)	3
Torgerson and Elbourne (2002)	6
Torgerson and Zhu (2003)	11
Torgerson *et al* (2002)	7
Total	**56**

It should be noted in passing that in several cases a study coded as a 'randomized trial' in the original meta-analysis was excluded from my analysis if, on examination of the original paper, it was discovered that the effect size had not been derived from a randomized comparison. As well as the threat to their validity from publication bias (i.e., not including relevant trials), the conclusions of systematic reviews may also be threatened by the inclusion of inappropriate studies. As noted previously, the RCT is the 'gold-standard' method for deriving causal inferences. During the process of undertaking systematic reviews in literacy research I have noted on many occasions that it can be difficult to determine whether or not a study actually *is* a RCT or a controlled trial. This was discussed to some extent in Item 1 of this portfolio. All the non-randomized trials which previous reviewers had included inappropriately, but which I excluded, are listed below.

Mathes and Fuchs (1994) undertook a systematic review of the efficacy of peer tutoring in reading with 'students with mild disabilities'. As part of their exhaustive search for potentially relevant studies, the authors contacted researchers known to have been involved in peer tutoring research to ask them for any unpublished reports or manuscripts that could be relevant to the review. A total of 11 studies were included in the review (Table 3, pp.68-9). The pooled effect size across all studies was 0.36 ($p<0.01$), which is an educationally and statistically significant result. The authors

concluded that 'peer tutoring in reading with students with mild disabilities can be effective' (p.76). The students with mild disabilities generally made greater progress in reading if they participated in peer tutoring reading interventions than control students who participated in typical teacher-directed reading instruction (without researcher direction), although the authors noted that the effect sizes were very variable and could have depended on the quality of the interventions being implemented.

This review included 11 experimental studies. It seems from Table 3 (pp.68-9) that nine of these trials used random allocation to intervention or control group (Carlton *et al*, 1985; McCracken, 1979; Russell and Ford, 1983; Scruggs and Osguthorpe, 1986, Experiment 1; Simmons, 1994; Simmons, 1995; Sindelar, 1982; Top and Osguthorpe, 1985; Top and Osguthorpe, 1987), and two of the studies used 'matching' without randomization and were, therefore, quasi-experiments (Lamport, 1982; Scruggs and Osguthorpe, 1986, Experiment 2). However, scrutiny of the original papers reveals that four of the other nine trials are not actually randomized controlled trials:

- Simmons (1995) did include a randomized trial, but the effect size reported in the meta-analysis had been calculated from a non-randomized comparison.
- Mathes and Fuchs (1994) described a second trial (Scruggs and Osguthorpe, 1986) thus: 'subjects randomly assigned to groups' (p.68). However, scrutiny of the original paper reveals that the pupils were selected for experimental or control groups by the participating teachers (p.188).
- A third trial (Top and Osguthorpe, 1987) was excluded because it employed a mixture of cluster and individual randomization and did not adjust for this in the analysis.

- A fourth trial (McCracken, 1979) described allocation thus: 'classes randomly assigned to treatment groups' (p.68). However, allocation is not described in the original paper and it is not evident from the paper that this study is actually a randomized controlled trial.

A fifth trial in the Mathes and Fuchs (1994) meta-analysis was excluded because assignment to treatment or control groups was not randomized (Simmons, 1994). Within the intervention group of 23 teachers, assignment to one of four peer tutoring conditions was random, but the remaining eight teachers assigned to the control condition (normal practice) had not been randomized. One trial retrieved from the Ehri *et al* (2001b) review was excluded due to huge attrition (Mantzicopoulos *et al*, 1992). The problems with this trial have been discussed in detail in Chapter 4. One trial described as randomized in the Bus and van IJzendoorn (1999) review was excluded because the method of allocation was not stated in the paper (Content *et al*, 1982), and therefore I could not be sure that this was in fact a randomized controlled trial.

One further trial was excluded because the original paper was unobtainable (Top and Osguthorpe, 1985).

The reviews were de-duplicated in a hierarchical progression, starting from the top of the table and working downwards. Therefore trials already retrieved from one systematic review higher up the table may also be present in a review further down the table. For example, four trials identified in Torgerson and Elbourne (2002) were also present in Torgerson and Zhu (2003). The number identified refers to the number of studies (RCTs), not the number of reports or articles, i.e., more than one trial was present in a number of articles or reports.

Table 7.2 shows the following characteristics of the 56 included trials: sample size, unit of allocation (i.e. individual or cluster), whether or not follow-up was blind, and the attrition rate.

Table 7.2: Characteristics of the 56 included trials

Study reference	Sample size	Cluster (number of clusters) or individual	Blinded follow-up?	Total drop out rates (number)	Published?	Effect size, as reported by reviewers
Baker *et al* (2000)	127	Ind.	Y	43	Y	0.30
Ball (1997)	27	Ind.	NS	0	Y	0.65
Ball and Blachman (1991)	60	Ind.	NS	1	Y	0.72
Berninger *et al* (1998), Study 1	24	Ind.	NS	0	Y	-0.05
Berninger *et al* (1998), Study 2	24	Ind.	NS	0	Y	0.32
Brown and Felton (1990)	47 Effective sample size: 11	Clust. (6)	NS	6	Y	0.48
Byrne and Fielding-Barnsley (1991)	128	Ind.	NS	2	Y	1.76
Byrne and Fielding-Barnsley (1995)	120	Ind.	NS	5	Y	0.32
Carlton *et al* (1985)	136 Effective sample size: 25	Clust. (12)	NS	0	Y	0.38
Cunningham (1990)	28	Ind.	NS	0	Y	0.48
De La Paz and Graham (1997)	21	Ind.	NS	0	Y	0.65
Elliott *et al* (2000)	140 Effective sample size: 13	Clust. (6)	NS	0	Y	-0.08
Fox and Routh (1976)	40	Ind.	NS	0	Y	0.20
Fox and Routh (1984)	21	Ind.	NS	0	Y	1.19
Gittelman and Feingold (1983)	61	Ind.	NS	5	Y	0.53
Golden (1990	31	Ind.	NS	1	Y	0.12
Greaney *et al* (1997)	36	Ind.	Y	0	Y	0.37
Haskell *et al* (1992)	24	Ind.	NS	0	Y	0.14
Hatcher *et al* (1994)	64	Ind.	Y	1	Y	0.19
Heise *et al* (1991)	55	Ind.	NS	0	Y	0.49
Hohn *et al* (1983)	16	Ind.	NS	0	Y	0.00
Jaben (1983)	49	Ind.	NS	0	Y	1.38
Jaben (1987)	50	Ind.	NS	0	Y	1.38
Jinkerson and Baggett (1993)	20	Ind.	NS	0	Y	-0.02

Jones (1994)	20	Ind.	NS	0	Y	1.25
Leach and Siddall (1990)	20	Ind.	NS	0	Y	1.99
Lee (1980)	70	Ind.	NS	0	Y	0.06
Lin *et al* (1991), Study 1	48	Ind.	NS	0	Y	-0.165
Lin *et al* (1991), Study 2	45	Ind.	NS	0	Y	-0.45
Loenen (1989)	81	Ind.	NS	0	Y	-0.36
Lovett *et al* (1989)	121	Ind.	NS	0 (first battery of tests)	Y	0.07
Lovett and Steinbach (1997)	28	Ind.	NS	NS	Y	0.49
Lovett *et al* (1990)	36	Ind.	NS	NS	Y	0.16
Lovett *et al* (2000)	37	Ind.	NS	0	Y	0.6
MacArthur *et al* (1990)	44	Ind.	Y	0	Y	0.35
Martinussen and Kirby (1998)	26	Ind.	NS	2	Y	0.62
Matthew (1996)	74	Ind.	NS	0	Y	-0.32
McClurg and Kasakow (1989)	35	Ind.	NS	0	Y	1.15
Mitchell and Fox (2001)	72	Ind.	Y	0	Y	-0.60
Morris *et al* (1990)	60	Ind.	NS	0	Y	0.69
O'Connor *et al* (1993)	21	Ind.	NS	0	Y	0.88
Reinking and Rickman (1990)	60	Ind.	NS	0	Y	0.168
Rimm-Kaufman *et al* (1999)	42	Ind.	Y	0	Y	0.43
Russell and Ford (1983)_	32	Ind.	NS	0	Y	0.75
Sindelair (1982)	29 Effective sample size: 15	Clust. (9)	NS	0	Y	0.07
Swanson and Trahan (1992), Study 1	60	Ind.	NS	NS	Y	-0.267
Swanson and Trahan (1992), Study 2	60	Ind.	NS	NS	Y	0.639
Torgesen *et al* (1999)	90	Ind.	NS	NS	Y	0.33
Umbach *et al* (1989)	31	Ind.	NS	0	Y	1.08
Vadasy *et al* (1997)	40	Ind.	NS	0	Y	0.21
Vellutino and Scanlon (1987)	30	Ind.	NS	0	Y	0.79
Watson (1988)	16	Ind.	Y	0	Y	0.007
Weiner (1994)	36	Ind.	NS	0	Y	0.24
Weiss *et al* (1988)	17	Ind.	NS	1	N	-0.13
Whitehurst *et al* (1994)	207 Effective sample size: 30	Clust. (15)	Y	40	Y	0.79
Zhang *et al* (1995)	22	Ind.	NS	0	Y	2.74

NS = not stated

It was not possible to test whether there was a relationship between publication status and effect size because only one 'unpublished' study was included in the dataset (Weiss *et al*, 1988). Similarly, it was not possible to look at intention to teach analysis because only one study reported attrition and kept participants in their original groups (Torgesen *et al*, 1999). Therefore, only three of the five hypotheses could be tested.

Cluster trials versus individually randomized studies.

In table 7.3 I compare the effect sizes in the 5 cluster trials with the 51 individually randomized studies. As the table shows, the cluster trials tended to have on average a smaller effect size than that of individually randomized trials. The difference (0.15), however, was not statistically significant (95% CI of difference = -0.42 to 0.72, p=0.60).

Table 7.3: Comparison of mean effect sizes between cluster and individually randomized trials.

Type of trial	N	Mean effect size, as reported by reviewers	Std. Deviation
Individual trial	51	0.48	0.63
Cluster trial	5	0.33	0.34

Blinded versus unblinded follow-up

Eight trials reported that the post-tests were performed 'blind' to the group allocation. The average effect size was smaller in those trials compared with those that did not state whether or not blinded follow-up had been used. The mean difference in effect size (0.28), although quite large, was not statistically significant (95% CI of difference –0.17 to 0.73, p=0.24).

Table 7.4: Comparison of trials with blinded follow-up compared with those with unblinded follow-up.

Blinded follow-up?	N	Mean effect size, as reported by reviewers	Std. Deviation
Blinded follow-up	8	0.23	0.40
Not stated	48	0.51	0.63

Assessment of attrition

There was a very small (0.02) (close to random) correlation between effect size and drop-out (i.e., the larger the drop-out rate the bigger the effect size). However, this tiny correlation was not statistically significant (p=0.91).

Meta-regression analysis

To assess whether any of the factors independently predicted the effect size after controlling for the others, a meta-regression was undertaken. Table 7.5 presents the results. As the table shows, none of the characteristics was statistically significant. However, blinded follow-up had the biggest effect size, and this was in the expected direction, and close to conventional levels of statistical significance.

Table 7.5: Weighted regression of effect size

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
(Constant)	.45	.44		1.01	.32	-.44	1.33
Cluster	.09	.44	.030	.20	.84	-.80	.98
Blinded follow-up	-.49	.26	-.31	-1.84	.07	-1.02	.046
Total drop out rates (number)	.01	.01	.15	.90	.37	-.01	.029

Discussion

In this chapter all the randomized trials identified in the tertiary review were retrieved and examined for their methodological quality. Interestingly, on closer inspection of studies within the seven systematic reviews it was found that six studies described in three meta-analyses as RCTs were actually not RCTs after all, again supporting the notion that, for really important policy issues, systematic reviews should be replicated by two independent review teams. It would require contact with the authors of these trials (some of them undertaken over 30 years ago) in order to verify their status as RCTs.

It was not possible to look in detail at key quality issues, as most trials simply did not report important aspects of trial quality, such as concealed allocation. Indeed, most trials failed to report other crucial aspects of trial quality such as blinded follow-up. The meta-regression did not find strong evidence of an association between effect size and cluster randomization or attrition. However, there was a reasonably large difference in effect size between studies that used blinded follow-up and those that did not state whether or not follow-up was blind. This finding supports the view that unblinded follow-up can lead to bias, and all post-tests, if at all possible, should be conducted and assessed by teachers or researchers who are not aware of the group status of the participants.

Conclusions

Trials in literacy learning are, in general, poorly reported. Most do not report the key aspects of a study design and conduct. Therefore, it is difficult to quality appraise such

trials. Because quality appraisal is so difficult, it is problematic to rely too heavily on current evidence to support different approaches to literacy learning. This whole field requires large, rigorous trials to be conducted as a matter of urgency.

References

Chapter 1

Cook, T.D. and Campbell, D.T. (1979) *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Boston: Houghton Mifflin Company.

Kjaergard, L.L., Villumsen, J. and Gluud, C. (2001) Reported methodological quality and discrepancies between large and small randomized trials in meta-analyses, *Annals of Internal Medicine*, 135(11): 982-989.

Schulz, K.F., Chalmers, I., Hayes, R.J. and Altman, D.G. (1995) Empirical evidence of bias: Dimensions of methodological quality associated with estimates of effects in controlled trials, *Journal of the American Medical Association*, 273(5): 408-12.

Troia, G.A. (1999) Phonological awareness intervention research: A critical review of the experimental methodology, *Reading Research Quarterly*, 34(1): 28-52.

Chapter 2

Begg, C.B. and Berlin, J.A. (1988) Publication bias: A problem in interpreting medical data, *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 151(3): 419-463.

Blok, H. (1999) Reading to young children in educational settings: A meta-analysis of recent research, *Language Learning*, 49(2): 343-371.

Cohen, J. (1962) The statistical power of abnormal-social psychology research: A review, *Journal of Abnormal and Social Psychology*, 65(3): 145-153.

Dear, K.B.G. and Begg, C.B. (1992) An approach for assessing publication bias prior to performing a meta-analysis, *Statistical Science*, 7(2): 237-245.

Dickersin, K., Chan, S., Chalmers, T.C., Sacks H.S. and Smith, H. (1987) Publication bias and clinical trials, *Controlled Clinical Trials*, 8: 343-353.

Dickersin, K. (1997) How important is publication bias? A synthesis of available data, *AIDS Education and Prevention*, 9 (Supplement A): 15-21.

Dickersin, K. (2002) *Reducing reporting biases*, in Chalmers, I., Milne, I., Trohler, U. (eds.) The James Lind Library (www.jameslindlibrary.org).

Durlak, J.A. and Lipsey, M.W. (1991) A practitioner's guide to meta-analysis, *American Journal of Community Psychology*, 19(3): 291-332.

Egger, M., Juni, P., Bartlett, C., Holenstein, F. and Sterne, J. (2003) How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? *Empirical Study Health Technology Assessment*, 7(1): 1-76.

Ehri, L.C, Nunes, S.R., Stahl, S.A. and Willows, D.M., (2001b) Systematic phonics instruction helps students learn to read: evidence from the national reading panel's meta-analysis, *Review of Educational Research*, 71: 393-447.

Evans, S. (1996) Statistician's comment, *British Medical Journal*, 312:125.

Fitz-Gibbon, C. (2004) Editorial: The need for randomized trials in social research, *Journal of Royal Statistical Society*, 167(1): 1-4.

Greenwald, A.G. (1975) Consequences of prejudice against the null hypothesis, *Psychological Bulletin*, 82(1): 1-20.

Hedges, L.V. and Olkin, I. (1980) Vote counting methods in research synthesis, *Psychological Bulletin*, 88(2): 359-369.

Hsu, L.M. (2002) Fail-safe Ns for one- versus two-tailed tests lead to different conclusions about publication bias, *Understanding Statistics*, 1(2): 85-100.

Iyengar, S. and Greenhouse, J.B. (1988) Selection models and the file drawer problem, *Statistical Science*, 3(1): 109-117.

Kulik, J.A. and Kulik, C-L.C. (1989) Meta-analysis in education, *International Journal of Educational Research*, 13: 221-340.

Light, R.J. and Pillemer, D.B. (1984) *Summing up: The science of reviewing research*, Cambridge, MA: Harvard University Press.

Lipsey, M.W., Crosse, S., Dunkle, J., Pollard, J. and Stobart, G. (1985) Evaluation: The state of the art and the sorry state of the science, *New Directions for Program Evaluation*, 27.

Lipsey, M.W. and Wilson, D.B. (1993) The efficacy of psychological, educational and behavioral treatment: Confirmation from meta-analysis, *American Psychologist*, 12: 1181-1209

Norris, J.M. and Ortega, L. (2000) Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis, *Language Learning*, 50(3): 417-528.

Rosenthal, R. (1978) Combining results of independent studies, *Psychological Bulletin*, 85: 185-193.

Rosenthal, R. (1979) The 'file drawer problem' and tolerance for null results, *Psychological Bulletin*, 86(3): 638-641.

Smart, R.G. (1964) The importance of negative results in psychological research, *Canadian Psychologist*, 5: 225-32.

Smith, M.L. (1980) Publication bias and meta-analysis, *Evaluation in Education*, 4: 22-24.

Sterling, T.D. (1959) Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa, *Journal of American Statistical Association*, 54: 30-34.

Sterling, T.D., Rosenbaum, W.L. and Weinkam, J.J. (1995) Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa, *The American Statistician*, 49(1): 108-112.

Sterne, J.A.C., Gavaghan, D. and Egger, M. (2000) Publication and related bias in meta-analysis, *Journal of Clinical Epidemiology*, 53(11): 1119-1129.

Sterne, J.A.C., Egger, M. and Smith, G.D. (2001) Investigating and dealing with publication and other biases, in Egger, M. Smith, G.D. and Altman, D.G. (eds.) *Systematic Reviews in Healthcare: Meta-Analysis in Context*, second edition, BMJ publishing group, London, 2001.

Sutton, A.J., Duval, S.J., Tweedie, R.L., Abrams, K.R.A. and Jones, D.R. (2000) Empirical assessment of the effect of publication bias on meta-analyses, *British Medical Journal*, 320: 1574-1577.

Torgerson, C.J. (2003) *Systematic reviews*, London: Continuum Books.

Torgerson, C.J., Porthouse, J. and Brooks, G. (2003) A systematic review and meta-analysis of randomised controlled trials evaluating interventions in adult literacy and numeracy, *Journal of Research in Reading*, 26(3): 234-255.

Truscott, J. (2004) The effectiveness of grammar instruction: Analysis of a meta-analysis, *English Teaching and Learning*, 28(3): 17-29.

Wang, M.C. and Bushman, B.J. (1998) Using the normal quantile plot to explore meta-analytic data sets, *Psychological Methods*, 3(1): 46-54.

Wilson, D.B. and Lipsey, M.W. (2001) The role of method in treatment effectiveness research: Evidence from meta-analysis, *Psychological Methods*, 6(4): 413-429.

Chapter 3

Bangert-Drowns, R.L. (1993) The word processor as an instructional tool: A meta-analysis of word processing in writing instruction, *Review of Educational Research,* 63(1): 69-93.

Blok, H. (1999) Reading to young children in educational settings: A meta-analysis of recent research, *Language Learning,* 49(2): 343-371.

Bus, A.G., van IJzendoorn, M.H. and Pellegrini, A.D. (1995) Joint book reading makes for success in learning to read: A meta-analysis on intergenerational transmission of literacy, *Review of Educational Research,* 65(1): 1-21.

Bus, A.G. and van IJzendoorn, M.H. (1999) Phonological awareness and early reading: A meta-analysis of experimental training studies, *Journal of Educational Psychology,* 91(3): 403-414.

Dubben, H.H. and Beck-Bornholdt, H.P. (2005) *British Medical Journal,* doi: 10.1136/bmj.38478.497164.F7 (published 3 June 2005).

Ehri, L.C., Nunes, S.R., Willows, D.M., Schuster, B.V., Yaghoub-Zadeh, Z. and Shanahan, T. (2001a) Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis, *Reading Research Quarterly,* 36(3): 250-287.

Ehri, L.C., Nunes, S.R., Stahl, S.A. and Willows, D.M. (2001b) Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis, *Review of Educational Research,* 71(3): 393-447.

Elbaum, B. and Vaughn, S. (2000) How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research, *Journal of Educational Psychology,* 92(4): 605-619.

Gersten, R. and Baker, S. (2001) Teaching expressive writing to students with learning disabilities: A meta-analysis, *Elementary School Journal,* 101(3): 251-272.

Haller, E.P., Child, D.A. and Walberg, H.J. (1988) Can comprehension be taught? A quantitative synthesis of 'meta-cognitive' studies, *Educational Researcher,* 17(9): 5-8.

Jeynes, W.H. and Littell, S.W. (2000) A meta-analysis of studies examining the effect of whole language instruction on the literacy of low-SES students, *Elementary School Journal,* 101(1): 21-33.

Mathes, P.G. and Fuchs, L.S. (1994) The efficacy of peer tutoring in reading for students with mild disabilities: A best evidence synthesis, *School Psychology Review,* 23(1): 59-80.

McAulay, L., Pham, Ba', Tugwell, P. and Moher, D. (2000) Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? *Lancet,* 356: 1228-1231.

Moher, D., Cook, D.J., Eastwood, S., Olkin, I., Rennie, D, Stroup, D.F. (1999) Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. Quality of Reporting of Meta-analyses, *Lancet,* 354: 1896-1900.

Terrin, N., Schmid, C.H. and Lau, J. (2005) In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias, *Journal of Clinical Epidemiology,* 58: 894-901.

Torgerson, C.J. and Elbourne, D. (2002) A systematic review and meta-analysis of the effectiveness of information and communication technology (ICT) on the teaching of spelling, *Journal of Research in Reading,* 25(2): 129-143.

Torgerson, C.J., King, S.E. and Sowden, A.J. (2002) Do volunteers in schools help children learn to read? A systematic review of randomized controlled trials, *Educational Studies,* 28(4): 433-444.

Torgerson, C.J. and Zhu, D. (2003) A systematic review and meta-analysis of the effectiveness of ICT on literacy learning in English, 5-16, in *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Berninger, V.W., Vermeulen, K., Abbott, R.D., McCutchen, D., Cotton, S., Cude, J., Dorn, S. and Sharon, T. (2003) Comparison of three approaches to supplementary reading instruction for low-achieving 2nd grade readers, *Language, Speech and Hearing Services in Schools*, 34(2): 101-116.

Blachman, B.A., Schatschneider, C., Fletcher, J.M., Francis, D.J., Clonan, S.M., Shaywitz, B.A. and Shaywitz, S.E. (2004) Effects of intensive reading remediation for second and third graders and a 1-year follow-up, *Journal of Educational Psychology*, 96(3): 444-461.

Brooks, G., Miles, J., Torgerson, C.J. and Torgerson, D.J. (2005) *A randomised trial of computer software in education, using CONSORT guidelines*, oral presentation at the 9th Social and Health Sciences Methodology Conference, Granada, Spain, September 2005.

Brown, I.S. and Felton, R.H. (1990) Effects of instruction on beginning reading skills in children at risk for reading disability, *Reading and Writing: An Interdisciplinary Journal*, 2(3): 223-241.

Camilli, G., Vargas, S. and Yurecko, M. (2003). '*Teaching Children to Read*: the fragile link between science and federal education policy.' *Education Policy Analysis Archives*, 11, no.15, retrieved 8 June 2005 http://epaa.asu.edu/epaa/v11n15/

Egger, M., Davey Smith, G., Schneider, M and Minder, C. (1997) Bias in meta-analysis detected by a simple graphical test, *BMJ*, 315: 629-634(13 September).

Ehri, L.C., Nunes, S.R., Stahl, S.A. and Willows, D.M. (2001b) Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis, *Review of Educational Research*, 71(3): 393-447.

Ehri, L. and Stahl, S. (2001) Beyond the smoke and mirrors: Putting out the fire, *Phi Delta Kappan*, 83(1): 17-20.

Fayne, H.R. and Bryant, N.D. (1981) Relative effects of various word synthesis strategies on the phonics achievement of learning disabled youngsters, *Journal of Educational Psychology*, 73(5): 616-623.

Garan (2001a) What does the report of the National Reading Panel really tell us about teaching phonics?, *Language Arts*, 79 (1): 61-71.

Garan (2001b) Beyond the smoke and mirrors: A critique of the National Reading Panel report on phonics, *Phi Delta Kappan*, 82(7): 500-506.

Garan (2001c) More smoking guns: A response to Linnea Ehri and Steven Stahl, *Phi Delta Kappan*, 83(1): 21-27.

Gittelman, R. and Feingold, I. (1983) Children with reading disorders –1. Efficacy of reading remediation, *Journal of Child Psychology and Psychiatry*, 24(2): 167-191.

Greaney, K.T., Tunmer, W.E. and Chapman, J.W. (1997) Effects of rime-based orthographic analogy training on the word recognition skills of children with reading disability, *Journal of Educational Psychology*, 89(4): 645-651.

Haskell, D.W., Foorman, B.R. and Swank, P. (1992) Effects of three orthographic/phonological units on first-grade reading, *Remedial and Special Education*, 13(2): 40-49.

Hatcher, P.J., Hulme, C. and Snowling, M.J. (2004). 'Explicit phoneme training combined with phonic reading instruction helps young children at risk of reading failure.' *Journal of Child Psychology and Psychiatry*, 45, 2, 338-58.

Herrera, J.A., Logan, D., Cooker, R., Morris, D. and Lyman, D. (1997) Phonological awareness and phonetic-graphic conversion: A study of the effects of two intervention paradigms with learning disabled children. Learning disability or learning difference?, *Reading Improvement*, 34(2): 71-89.

Jenkins, J.R., Peyton, J.A., Sanders, E.A. and Vadasy, P.F. (2004) Effects of reading decodable texts in supplemental first-grade tutoring, *Scientific Studies of Reading*, 8(1): 53-85.

Johnston, R.S. and Watson, J.E. (2004) Accelerating the development of reading, spelling and phonemic awareness skills in initial readers, *Reading and Writing: An Interdisciplinary Journal*, 17(4): 327-357.

Leach, D.J. and Siddall, S.W. (1990) Parental involvement in the teaching of reading: A comparison of hearing reading, paired reading, pause, prompt, praise and direct instruction methods, *British Journal of Educational Psychology*, 60(3): 349-355.

Lovett, M.W., Ransby, M.R., Hardwick, N., Johns, M.S. and Donaldson, S.A. (1989) Can dyslexia be treated? Treatment-specific and generalized treatment effects in dyslexic children's response to remediation, *Brain and Language*, 37(1): 90-121.

Lovett, M.W., Warren-Chaplin, P.M., Ransby, M.J. and Borden, S.L. (1990) Training the word recognition skills of reading disabled children: Treatment and transfer effects, *Journal of Educational Psychology*, 82(4): 769-780.

Lovett, M.W. and Steinbach, K.A. (1997) The effectiveness of remedial programs for reading disabled children of different ages: Does the benefit decrease for older children?, *Learning Disability Quarterly*, 20(3): 189-210.

Lovett, M.W., Lacerenza, L., Borden, S.L., Frijters, J.C., Steinbach, K.A. and De Palma, M. (2000) Components of effective remediation for developmental reading disabilities: Combining phonological and strategy-based instruction to improve outcomes, *Journal of Educational Psychology*, 92(2): 263-283.

Manzticopoulos, P., Morrison, D., Stone, E. and Setrakian, W. (1992) Use of the SEARCH/TEACH tutoring approach with middle-class students at risk for reading failure, *The Elementary School Journal*, 92(5): 573-586.

Martinussen, R.L. and Kirby, J.R. (1998) Instruction in successive phonological processing to improve the reading acquisition skills of at-risk kindergarten children, Developmental Disabilities Bulletin, 26(2): 19-39.

O'Connor, R.E. and Padeliadu, S. (2000) Blending versus whole word approaches in first grade remedial reading: Short-term and delayed effects on reading and spelling words, *Reading and Writing: An Interdisciplinary Journal*, 13(1-2): 159-182.

Oudeans, M.K. (2003) Integration of letter-sound correspondences and phonological awareness skills of blending and segmenting: A pilot study examining the effects of instructional sequence on word reading for kindergarten children with low phonological awareness, *Learning Disability Quarterly*, 26(4): 258-280.

Skailand, D.B. (1971) *A comparison of four language units in teaching beginning reading*, Paper presented at the meeting of the American Educational Research Association, New York, USA, 4-7 February 1971.

Sullivan, H.J, Okada, M. and Niedermeyer, F.C. (1971) Learning and transfer under two methods of word-attack instruction, *American Educational Research Journal*, 8 (2): 227-239.

Torgerson, C.J. (2003) *Systematic Reviews*, London: Continuum Books.

Torgesen, J.K., Wagner, R.K., Lindamood, P., Rose, E., Conway, T. and Gravan, C. (1999) Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction, *Journal of Educational Psychology*, 91(4): 579-593.

Torgesen, J.K., Alexander, A.W., Wagner, R.K., Rashotte, C.A., Voeller, K.K.S. and Conway, T. (2001) Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches, *Journal of Learning Disabilities*, 34(1): 33-58.

Umbach, B. and Halpin, G. (1989) Teaching reading to low performing first graders in rural schools: A comparison of two instructional approaches, *Journal of Instructional Psychology*, 16(3): 112-121.

Vadasy, P.F., Jenkins, J.R., Antil, L.R., Wayne, S.K. and O'Connor, R.E. (1997) The effectiveness of one-to-one tutoring by community tutors for at-risk beginning readers, *Learning Disability Quarterly,* 20(2): 126-139.

Walton, P.D., Walton, L.M. and Felton, K. (2001) Teaching rime analogy or letter recoding reading strategies to pre-readers: Effects on pre-reading skill and word reading, *Journal of Educational Psychology*, 93(1): 160-180.

Walton, P.D. and Walton, L.M. (2002) Beginning reading by teaching in rime analogy: Effects on phonological skills, letter-sound knowledge, working memory, and word-reading strategies, *Scientific Studies of Reading*, 6(1): 79-115.

Chapter 5

Altman, D.G. and Dore, C.J. (1990) Randomisation and baseline comparisons in clinical trials, *Lancet*, 335: 149-153.

Altman, D.G., Schulz, K.F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., Gotzsche, P.C. and Lang, T. (2001) The revised CONSORT statement for reporting randomized trials: Explanation and elaboration, *Annals of Internal Medicine*, 134(8): 663-694

Begg, C.B. and Berlin, J.A. (1988) Publication bias: A problem in interpreting medical data, *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 151(3): 419-463.

Begg, C.B., Cho, M.K., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D., Schulz, K.F., Simel, D. and Stroup, D.F. (1996) Improving the quality of reporting of randomized controlled trials: The CONSORT statement, *Journal of the American Medical Association*, 276(8): 637-639.

Boruch, R.F., McSweeney, A.J. and Soderstrom, E.J. (1978) Randomized field experiments for program planning, development and evaluation: An illustrative bibliography, *Evaluation Quarterly*, 2(4): 655-695.

Boruch, R.F. (1994) The future of controlled randomized experiments: A briefing, *Evaluation Practice*, 15(3): 265-74.

Chalmers, I. (2001) Comparing like with like: Some historical milestones in the evolution of methods to create unbiased comparison groups in therapeutic experiments, *International Journal of Epidemiology*, 30(5): 1156-1164.

Cook, T.D. and Campbell, D.T. (1979) *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Boston: Houghton Mifflin Company.

Craven, R.G., Marsh, H.W., Debus, R.L. and Jayasinghe, U. (2001) Diffusion effects: Control group contamination threats to the validity of teacher-administered interventions, *Journal of Educational Psychology*, 93(3): 639-645.

Glass, G.V. (1977) Integrating findings: The meta-analysis of research, in Shulman, L. (ed.), *Review of Research in Education*, 5: 351-379, Itasca, Il: Peacock.

Glass, G.V., McGaw, B. and Smith, M.L. (1981) *Meta-Analysis in Social Research*, Beverly Hills, CA: Sage.

Hannon, P. (2000) *Reflecting on Literacy in Education*, London: RoutledgeFalmer.

Johnston, R.S. and Watson, J.E. (2004) Accelerating the development of reading, spelling and phonemic awareness skills in initial readers, *Reading and Writing: An Interdisciplinary Journal*, 17: 327-357.

Kjaergard, L.L., Villumsen, J. and Gluud, C. (2001) Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses, *Annals of Internal Medicine*, 135(11): 982-89.

Kulik, J.A. and Kulik, C.-L.C. (1989) Meta-analysis in education, *International Journal of Educational Research*, 13: 221-340.

Lindquist, E.F. (1940) *Statistical Analysis in Educational Research*, Boston: Houghton Mifflin Company.

Lipsey, M.W., Crosse, S., Dunkle, J., Pollard, J and Stobart, G. (1985) Evaluation: the state of the art and the sorry state of the science, *New Directions for Program Evaluation*, 27: 7-28.

Lipsey, M.W. and Wilson, D.B. (1993) The efficacy of psychological, educational and behavioral treatment: Confirmation from meta-analysis, *American Psychologist*, 48(12): 1181-1209.

Moher, D., Dulberg, C. S. and Wells, G. A. (1994) Statistical power, sample size, and their reporting in randomized controlled trials, *Journal of the American Medical Association*, 272(2): 122-24.

Oakley, A. (2000) *Experiments in Knowing: Gender and Method in the Social Sciences*, Cambridge: Polity Press.

Schulz, K.F., Chalmers, I., Hayes, R.J. and Altman, D.G. (1995) Empirical evidence of bias: Dimensions of methodological quality associated with estimates of effects in controlled trials, *Journal of the American Medical Association*, 273(5): 408-12.

Schulz, K.F. and Grimes, D.A. (2002) Allocation concealment in randomised trials: Defending against deciphering, *Lancet*, 359: 614-618.

Shadish, W.R., Cook, T.D. and Campbell, D.T. (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Boston: Houghton Mifflin Company.

Slavin, R.F. (1984) Meta-analysis in education: How has it been used? *Educational Researcher*, 13(8): 6-15.

Torgerson, C.J. and Torgerson, D.J. (2001) The need for randomised controlled trials in educational research, *British Journal of Educational Studies*, 49(3): 316-28.

Torgerson, C.J. and Elbourne, D. (2002) A systematic review and meta-analysis of the effectiveness of information and communication technology (ICT) on the teaching of spelling, *Journal of Research in Reading*, 25(2): 129-143.

Torgerson, C.J., King, S.E. and Sowden, A.J. (2002) Do volunteers in schools help children learn to read? A systematic review of randomised controlled trials, *Educational Studies*, 28(4): 433-444.

Torgerson, C.J., Porthouse, J. and Brooks, G. (2003) A systematic review and meta-analysis of randomised controlled trials evaluating interventions in adult literacy and numeracy, *Journal of Research in Reading*, 26(3): 234-255.

Torgerson, C.J. (2003) *Systematic Reviews*, London: Continuum Books.

Torgerson, C.J. and Zhu, D. (2003) A systematic review and meta-analysis of the effectiveness of ICT on literacy learning in English, 5-16, in *Research Evidence in Education Library*, London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Torgerson, C.J., Torgerson, D.J., Birks, Y.F. and Porthouse, J. (2005, in press) A comparison of the quality of randomised controlled trials in education and health, *British Educational Research Journal*.

Wilson, D.B. and Lipsey, M.W. (2001) The role of method in treatment effectiveness research: Evidence from meta-analysis, *Psychological Methods*, 6(4): 413-429.

Altman, D.G. and Dore, C.J. (1990) Randomisation and baseline comparisons in clinical trials, *Lancet*, 335: 149-153.

Altman, D.G. (1996) Better reporting of randomised controlled trials: The CONSORT statement, *British Medical Journal*, 313: 570-571.

Altman, D.G., Schulz, K.F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., Gotzsche, P.C. and Lang, T. (2001) The revised CONSORT statement for reporting randomized trials: Explanation and elaboration, *Annals of Internal Medicine*, 134(8): 663-694.

Alvermann, D.E. and Moore, D.W. (1991) Secondary school reading, in Barr, R., Kamil, M.L., Mosenthal, P. and Pearson, P.D. (eds.) *Handbook of Reading Research*, (Vol II, pp.951-983), White Plains, NY: Longman.

Ball, E.W. and Blachman, B.A. (1991) Does phoneme awareness training in kindergarten make a difference in early word recognition and developmental spelling? *Reading Research Quarterly*, 26(1): 49-66.

Begg, C.B., Cho, M.K., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D., Schulz, K.F., Simel, D. and Stroup, D.F. (1996) Improving the quality of reporting of randomized controlled trials: The CONSORT statement, *Journal of the American Medical Association*, 276(8): 637-639.

Bentin, S. and Leshem, H. (1993) On the interaction between phonological awareness and reading acquisition: It's a two-way street, *Annals of Dyslexia*, 43: 125-148

Boruch, R.F. *Randomized Experiments for Planning and Evaluation: A Practical Guide,* Applied Social Research Methods Series, Sage publications, London, 1997.

Byrne, B. and Fielding-Barnsley, R. (1991) Evaluation of a program to teach phonemic awareness to young children, *Journal of Educational Psychology*, 83(4): 451-455.

Byrne, B. and Fielding-Barnsley (1993) Evaluation of a program to teach phonemic awareness to young children: A 1-year follow-up, *Journal of Educational Psychology*, 85(1): 104-111.

Byrne, B. and Fielding-Barnsley, R. (1995) Evaluation of a program to teach phonemic awareness to young children: A 2- and 3-year follow-up and a new preschool trial, *Journal of Educational Psychology*, 87(3): 488-503.

Campbell, D.T. and Stanley, J.C. (1966) *Experimental and quasi-experimental designs for research*, Chicago: Rand McNally.

Castle, J., Riach, J. and Nicholson, T. (1994) Getting off to a better start in reading and spelling: The effects of phonemic awareness instruction within a whole language program, *Journal of Educational Psychology*, 86(3): 350-359.

Cook, T.D. and Campbell, D.T. (1979) *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Boston: Houghton Mifflin Company.

Cunningham, A. (1990) Explicit versus implicit instruction in phonemic awareness, *Journal of Experimental Child Psychology*, 50(3): 429-444.

Ehri, L.C., Nunes, S.R., Willows, D.M., Schuster, B.V., Yaghoub-Zadeh, S. and Shanahan, T. (2001a) Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis, *Reading Research Quarterly*, 36(3): 250-287.

Foster, K.C., Erickson, G.C., Foster, D.F., Brinkman, D. and Torgesen, J.K. (1994) Computer administered instruction in phonological awareness: Evaluation of the Daisyquest Program, *The Journal of Research and Development in Education*, 27(2): 126-137.

Freiman, J.A., Chalmers, T.C., Smith, H. and Kuebler, R.R. (1978) The importance of beta, the type II error, and sample size in the design and interpretation of the randomized controlled trial: Survey of 71 'negative' trials, *New England Journal of Medicine*, 299(13): 690-694.

Hatcher, P., Hulme, C. and Ellis, A. (1994) Ameliorating early reading failure by integrating the teaching of reading and phonological skills: The phonological linkage hypothesis, *Child Development*, 65(1): 41-57.

Hedges, L.V. (2000) Using converging evidence in policy formation: The case of class size research, *Evaluation and Research in Education*, 14(3&4): 193-205.

Hohn, W. E., and Ehri, L. C. (1983) Do alphabet letters help prereaders acquire phonemic segmentation skill? *Journal of Educational Psychology*, 75(5): 752-762.

Jadad, A.R., Moore, A., Carroll, D., Jenkinson, C., Reynolds, J.M., Gavaghan, D.J. and McQuay, H.J. (1996) Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials*, 17: 1-12.

Johnston, R.S. and Watson, J.E. (2004) Accelerating the development of reading, spelling and phonemic awareness skills in initial readers, *Reading and Writing: an Interdisciplinary Journal*, 17: 327-357.

Juni, P., Witschi, A., Bloch, R. and Egger, M. (1999) The hazards of scoring the quality of clinical trials for meta-analysis, *Journal of the American Medical Association*, 282(11): 1054-60.

Kjaergard L.L., Villumsen J. and Gluud C. (2001) Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses, *Annals of Internal Medicine*, 135(11): 982-89.

Kozminsky, L. and Kozminsky, E. (1995) The effects of early phonological awareness training on reading success, *Learning and Instruction*, 5(3): 187-201.

Kulik, J.A. and Kulik, C.-L.C. (1989) Meta-analysis in education, *International Journal of Educational Research*, 13(3): 221-340.

Lie, A. (1991) Effects of a training programme for stimulating skills in first-grade children, *Reading Research Quarterly*, 26(3): 234-250.

Lysynchuk, L.M., Pressley, M., d'Ally, H., Smith, M. and Cake, H. (1989) A methodological analysis of experimental studies of comprehension strategy instruction, *Reading Research Quarterly*, 24(4): 458-470.

Moher, D., Dulberg, C.S. and Wells, G.A. (1994) Statistical power, sample size, and their reporting in randomized controlled trials, *Journal of the American Medical Association*, 272(2): 122-24.

O'Connor, R.E., Jenkins, J.R., Leicester, N. and Slocum, T. A. (1993) Teaching phonological awareness to young children with learning disabilities, *Exceptional Children*, 59(6): 532-546.

O'Connor, R.E., Jenkins, J.R. and Slocum, T.A. (1995) Transfer among phonological tasks in kindergarten: Essential instructional content, *Journal of Educational Psychology*, 87(2): 202-217.

Ridgeway, V.G., Dunston, P.M. and Qian, G. (1993) A methodological analysis of teaching and learning strategy research at the secondary school level, *Reading Research Quarterly*, 28(4): 334-349.

Schulz, K.F., Chalmers, I., Hayes, R.J. and Altman, D.G. (1995) Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials, *Journal of the American Medical Association*, 273(5): 408-12.

Schulz, K.F. (1995) Subverting randomization in controlled trials, *Journal of the American Medical Association*, 274(18): 1456-8.

Schulz, K.F. and Grimes. D.A. (2002) Allocation concealment in randomised trials: Defending against deciphering, *Lancet*, 359: 614-18.

Schwartz, D. and Lellouch, D. (1967) Explanatory and pragmatic attitudes in therapeutic trials, *Journal of Chronic Diseases*, 20(8): 637-648.

Slocum, T.A., O'Connor, R.E. and Jenkins, J.R. (1993) Transfer among phonological manipulation skills, *Journal of Educational Psychology*, 85(4): 618-630.

Torgesen, J.K. and Davis, C. (1996) Individual difference variables that predict response to training in phonological awareness, *Journal of Experimental Child Psychology*, 63(1): 1-21.

Torgerson, C.J. and Elbourne D. (2002) A systematic review and meta-analysis of the effectiveness of information and communication technology (ICT) on the teaching of spelling, *Journal of Research in Reading*, 25(2): 129-43.

Torgerson, C.J. and Torgerson, D.J. (2001) The need for randomised controlled trials in educational research, *British Journal of Educational Studies*, 49(3): 316-328.

Torneus, M. (1984) Phonological awareness and reading: A chicken and egg problem? *Journal of Educational Psychology*, 76(6): 1346-1358.

Troia, G.A. (1999) Phonological awareness intervention research: A critical review of the experimental methodology, *Reading Research Quarterly*, 34(1): 28-52.

Uhry, J.K. and Shepherd, M.J. (1993) Segmentation/spelling instruction as part of a first-grade reading program: Effects on several measures of reading, *Reading Research Quarterly*, 28(3): 219-233.

Ukoumonne, O.C., Gulliford, M.C., Chinn, S., Sterne, J.A.C. and Burney, P.G.J. (1998) Evaluation of healthcare interventions at area and organization level, in Black, N., Brazier, R., Fitzpatrick, R. and Reeves, B. *Health Services Research Methods*, London: BMJ Publications.

Vellutino, F.R. and Scanlon, D.M. (1987) Phonological coding, phonological awareness, and reading ability: Evidence from a longitudinal and experimental study, *Merrill-Palmer Quarterly*, 33(3): 321-363.

Weiner, S. (1994) Effects of phonemic awareness training on low- and middle-achieving first graders' phonemic awareness and reading ability, *Journal of Reading Behavior*, 26(3): 277-300.

Baker, S., Gersten, R. and Keating, T. (2000) When less may be more: A 2-year longitudinal evaluation of a volunteer tutoring program requiring minimal training, *Reading Research Quarterly,* 35(4): 494-519.

Ball, E.W. (1997) Phonemic awareness: Implications for whole language and emergent literacy problems, *Topics in Language Disorders*, 17(3): 14-26.

Ball, E.W. and Blachman, B.A. (1991) Does phoneme awareness training in kindergarten make a difference in early word recognition and developmental spelling? *Reading Research Quarterly*, 26(1): 49-66.

Berninger, V., Abbot, R., Rogan, L., Reed, E., Abbot, S., Brooks, A., Vaughan, K. and Graham, S. (1998) Teaching spelling to children with specific learning disabilities: The mind's ear and eye beat the computer or pencil, *Learning Disability Quarterly*, 21(2): 106–122.

Brown, I.S. and Felton, R.H. (1990) Effects of instruction on beginning reading skills in children at risk for reading disability, *Reading and Writing: An Interdisciplinary Journal*, 2(3): 223-241.

Brooks, G., Miles, J., Torgerson, C.J. and Torgerson, D.J. (2005) A randomised trial of computer software in education, using CONSORT guidelines, oral presentation at the 9th Social and Health Sciences Methodology Conference, Granada, Spain, September 2005.

Bus, A.G. and van IJzendoorn, M.H. (1999) Phonological awareness and early reading: A meta-analysis of experimental training studies, *Journal of Educational Psychology,* 91(3): 403-414.

Byrne, B. and Fielding-Barnsley, R. (1991) Evaluation of a program to teach phonemic awareness to young children, *Journal of Educational Psychology*, 83(4): 451-455.

Byrne, B. and Fielding-Barnsley, R. (1995) Evaluation of a program to teach phonemic awareness to young children: A 2- and 3-year follow-up and a new preschool trial, *Journal of Educational Psychology*, 87(3): 488-503.

Carlton, M.B., Litton, F.W. and Zinkgraf, S.A. (1985) The effects of an intraclass peer-tutoring program on the sight-word recognition of students who are mildly mentally retarded, *Mental Retardation*, 23(2): 74-78.

Content, A., Morais, J., Alegria, J. and Bertelson, P. (1982) Accelerating the development of phonetic segmentation skills in kindergartners, *Cahiers de Psychologie Cognitive*, 2(3): 259-269.

Cunningham, A. (1990) Explicit versus implicit instruction in phonemic awareness, *Journal of Experimental Child Psychology*, 50(3): 429-444.

De La Paz, S. and Graham, S. (1997) Effects of dictations and advanced planning instruction on the composing of students with writing and learning problems, *Journal of Educational Psychology*, 89(2): 203-222.

Ehri, L.C., Nunes, S.R., Stahl, S.A. and Willows, D.M. (2001b) Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis, *Review of Educational Research*, 71(3): 393-447.

Elliott, J., Arthurs, J. and Williams, R. (2000) Volunteer support in the primary classroom: The long-term impact of one initiative upon children's reading performance, *British Educational Research Journal,* 26(2): 227-244.

Fox, B. and Routh, D.K. (1976) Phonemic analysis and synthesis as word-attack skills, *Journal of Educational Psychology*, 68(1): 70-74.

Fox, B and Routh, D.K. (1984) Phonemic analysis and synthesis as word-attack skills: Revisited, *Journal of Educational Psychology*, 76(6): 1059-1064.

Gersten, R. and Baker, S. (2001) Teaching expressive writing to students with learning

disabilities: A meta-analysis, *Elementary School Journal*, 101(3): 251-272.
Gittelman, R. and Feingold, I. (1983) Children with reading disorders –1. Efficacy of reading remediation, *Journal of Child Psychology and Psychiatry*, 24(2): 167-191.
Golden, N., Gersten, R. and Woodward, J. (1990) Effectiveness of guided practice during remedial reading instruction: An application of computer-managed instruction, *Elementary School Journal,* 90(3): 291-304.
Greaney, K.T., Tunmer, W.E. and Chapman, J.W. (1997) Effects of rime-based orthographic analogy training on the word recognition skills of children with reading disability, *Journal of Educational Psychology*, 89(4): 645-651.
Haskell, D.W., Foorman, B.R. and Swank, P.R. (1992) Effects of three orthographic/phonological units on first-grade reading, *Remedial and Special Education*, 13(2): 40-49.
Hatcher, P., Hulme, C. and Ellis, A. (1994) Ameliorating early reading failure by integrating the teaching of reading and phonological skills: The phonological linkage hypothesis, *Child Development*, 65(1): 41-57.
Heise, B. L., Papalewis, R. and Tanner, D.E. (1991) Building base vocabulary with computer-assisted instruction, *Teacher Education Quarterly*, 18(1): 55-63.
Hohn, W.E., and Ehri, L.C. (1983) Do alphabet letters help prereaders acquire phonemic segmentation skill? *Journal of Educational Psychology*, 75(5): 752-762.
Jaben, T.H. (1983) The effects of creativity training on learning disabled students' creative written expression, *Journal of Learning Disabilities*, 16(5): 264-265.
Jaben, T.H. (1987) Effects of training on learning disabled students' creative written expression, *Psychological Reports*, 60(1): 23-26.
Jinkerson, L. and Baggett, P. (1993) Spell checkers: Aids in identifying and correcting spelling errors, *Journal of Computing in Childhood Education*, 4(3-4): 291–306.
Jones, I. (1994) The effect of a word processor on the written composition of pupils, *Computers in the Schools,* 11(2): 43-54.
Kjaergard L.L., Villumsen J. and Gluud C. (2001) Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses, *Annals of Internal Medicine*, 135(11): 982-89.
Lamport, K.C. (1982) The effect of inverse tutoring on reading disabled students in public school settings, *Dissertation Abstracts International*, 44: 729 (University Microfilms No.: 83-15, 707).
Leach, D.J. and Siddell, S.W. (1990) Parental involvement in the teaching of reading: A comparison of hearing reading, paired reading, pause, prompt, praise and direct instruction methods, *British Journal of Educational Psychology*, 60(3): 349-355.
Lee, V.E., Brooks-Gunn, J., Schnur, E. and Liaw, F.R. (1990) Are Head Start effects sustained? A longitudinal follow-up comparison of disadvantaged children attending Head Start, no preschool, and other preschool programs, *Child Development,* 61(2): 495-507.
Lin, A., Podell, D.M. and Rein, N. (1991) The effects of CAI on word recognition in mildly mentally handicapped and non-handicapped learners, *Journal of Special Education Technology,* 11(1): 16-25.
Loenen, A. (1989) The effectiveness of volunteer reading help and the nature of the reading help provided in practice, *British Educational Research Journal,* 15(3): 297-316.
Lovett, M.W., Ransby, M.R., Hardwick, N., Johns, M.S. and Donaldson, S.A. (1989) Can dyslexia be treated? Treatment-specific and generalized treatment effects in dyslexic children's response to remediation, *Brain and Language*, 37(1): 90-121.

Lovett, M.W., Warren-Chaplin, P.M., Ransby, M.J. and Borden, S.L. (1990) Training the word recognition skills of reading disabled children: Treatment and transfer effects, *Journal of Educational Psychology*, 82(4): 769-780.

Lovett, M.W. and Steinbach, K.A. (1997) The effectiveness of remedial programs for reading disabled children of different ages: Does the benefit decrease for older children? *Learning Disability Quarterly*, 20(3): 189-210.

Lovett, M.W., Lacerenza, L., Borden, S.L., Frijters, J.C., Steinbach, K.A. and De Palma, M. (2000) Components of effective remediation for developmental reading disabilities: Combining phonological and strategy-based instruction to improve outcomes, *Journal of Educational Psychology*, 92(2): 263-283.

MacArthur, C.A., Haynes, J.A., Malouf, D.B., Harris, K. and Owings, M. (1990) Computer assisted instruction with learning disabled students: Achievement, engagement, and other factors that influence achievement, *Journal of Educational Computing Research*, 6(3): 311-328.

Manzticopoulos, P., Morrison, D., Stone, E. and Setrakian, W. (1992) Use of the SEARCH/TEACH tutoring approach with middle-class students at risk for reading failure, *The Elementary School Journal*, 92(5): 573-586.

Martinussen, R.L. and Kirby, J.R. (1998) Instruction in successive phonological processing to improve the reading acquisition skills of at-risk kindergarten children, *Developmental Disabilities Bulleting*, 26(2): 19-39.

Mathes, P.G. and Fuchs, L.S. (1994) The efficacy of peer tutoring in reading for students with mild disabilities: A best evidence synthesis, *School Psychology Review*, 23(1): 59-80.

Matthew, K.I. (1996) The impact of CD-ROM storybooks on children's reading comprehension and reading attitude, *Journal of Educational Multimedia and Hypermedia*, 5(3-4): 379-94.

McClurg, P.A. and Kasakow, N. (1989) Wordprocessors, spelling checkers, and drill and practice programmes: Effective tools for spelling instruction? *Journal of Educational Computing Research*, 5(2): 187-198.

McCracken, S.J. (1979) The effect of a peer tutoring program utilizing data-based instruction on the word recognition and reading comprehension skills of secondary age level handicapped students, *Dissertation Abstracts International*, 40(08-A): 4516. (University Microfilms No. 79-24, 399).

Mitchell, M.J. and Fox, B.J. (2001) The effects of computer software for developing phonological awareness in low-progress readers, *Reading Research and Instruction*, 40(4): 315-332.

Morris, D., Shaw, B. and Perney, J. (1990) Helping low readers in grades 2 and 3: An after-school volunteer tutoring program, *The Elementary School Journal*, 91(2): 133-150.

O'Connor, R.E., Jenkins, J.R., Leicester, N. and Slocum, T. A. (1993) Teaching phonological awareness to young children with learning disabilities, *Exceptional Children*, 59(6): 532-546.

Reinking, D. and Rickman, S.S. (1990) The effects of computer-mediated texts on the vocabulary learning and comprehension of intermediate-grade readers, *Journal of Reading Behavior*, 22(4): 395-411.

Rimm-Kaufman, S.E., Kagan, J. and Byers, H. (1999) The effectiveness of adult volunteer tutoring on reading among "at risk" first grade children, *Reading Research and Instruction*, 38(2): 143-52.

Russell, T. and Ford, D.F. (1983) Effectiveness of peer tutors vs resource teachers, *Psychology in the Schools*, 20(4): 436-441.

Schulz, K.F., Chalmers, I., Hayes, R.J. and Altman, D.G. (1995) Empirical evidence of bias: Dimensions of methodological quality associated with estimates of

treatment effects in controlled trials, *Journal of the American Medical Association*, 273(5): 408-12.

Scruggs, T.E. and Osguthorpe, R.T. (1986) Tutoring interventions within social education settings: A comparison of cross-age and peer tutoring, *Psychology in the schools*, 23(2): 148-154.

Simmons, D.C., Fuchs, L.S., Pate, J. and Mathes, P.G. (1994), Importance of complexity and role reciprocity to classwide peer tutoring, *Learning Disabilities Research and Practice*, 9(4): 203-212.

Simmons, D.C., Fuchs, L.S., Fuchs, D., Mathes, P.G. and Pate, J. (1995), Effects of explicit teaching and peer-mediated instruction on the reading achievement of learning-disabled and low-performing students, *Elementary School Journal*, 95(5): 387-408 .

Sindelair, P.T. (1982) The effects of cross-age tutoring on the comprehension skills of remedial reading students, *The Journal of Special Education*, 16(2): 199-206.

Swanson, H.L. and Trahan, M.F. (1992) Learning disabled readers' comprehension of computer mediated text: The influence of working memory, metacognition and attribution, *Learning Disabilities Research and Practice*, 7(2): 74-86.

Top, B.L. and Osguthorpe, R.T. (1985) The effects of reverse role tutoring on reading achievement and self-concept, in R.T. Osguthorpe. W.D. Eiserman, L. Shisler, B.L. Top and T.E. Scruggs *Handicapped children as tutors: Final report (1984-1985)*. Document submitted to the Office of Special Education and Rehabilitation Services, U.S. Department of Education, Washington, DC (ERIC Document Reproductive Service No. ED 267 545).

Top, B.L. and Osguthorpe, R.T. (1987) Reverse role tutoring: The effects of handicapped students tutoring regular class students, *Elementary School Journal*, 87(4): 413-423.

Torgesen, J.K., Wagner, R.K., Lindamodd, P., Rose, E., Conway, T. and Gravan, C. (1999) Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction, *Journal of Educational Psychology*, 91(4): 579-593.

Torgerson, C.J. and Elbourne D. (2002) A systematic review and meta-analysis of the effectiveness of information and communication technology (ICT) on the teaching of spelling, *Journal of Research in Reading*, 25(2): 129-43.

Torgerson, C.J., King, S.E. and Sowden, A.J. (2002) Do volunteers in schools help children learn to read? A systematic review of randomized controlled trials, *Educational Studies*, 28(4): 433-444.

Torgerson, C.J. and Zhu, D. (2003) A systematic review and meta-analysis of the effectiveness of ICT on literacy learning in English, 5-16, in *Research Evidence in Education Library*, London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Troia, G.A. (1999) Phonological awareness intervention research: A critical review of the experimental methodology, *Reading Research Quarterly*, 34(1): 28-52.

Umbach, B., Darch, C. and Halpin, G. (1989) Teaching reading to low performing first graders in rural schools: A comparison of two instructional approaches, *Journal of Instructional Psychology*, 16(3): 112-121.

Vadasy, P.F., Jenkins, J.R., Antil, L.R., Wayne, S.K. and O'Connor, R.E. (1997) The effectiveness of one-to-one tutoring by community tutors for at-risk beginning readers, *Learning Disability Quarterly*, 20(2): 126-139.

Vellutino, F.R. and Scanlon, D.M. (1987) Phonological coding, phonological awareness, and reading ability: Evidence from a longitudinal and experimental study, *Merrill-Palmer Quarterly*, 33(3): 321-363.

Watson, A.J. (1988) Developmental spelling: A word categorizing instructional experiment, in *Journal of Educational Research*, 82(2): 82-88.

Weiner, S. (1994) Effects of phonemic awareness training on low- and middle-achieving first graders' phonemic awareness and reading ability, *Journal of Reading Behavior*, 26(3): 277-300.

Appendices

Appendix A

Carole J. Torgerson, Publication bias: The Achilles' heel of systematic reviews? *British Journal of Educational Studies*, in press, 54(1), 2006.

Appendix B

Carole J Torgerson, David J Torgerson, Yvonne F Birks, Jill Porthouse A Comparison of Randomised Controlled Trials in Health and Education, *British Educational Research Journal*, in press, 2005.

Appendix A

Carole J. Torgerson, Publication bias: The Achilles' heel of systematic reviews? *British Journal of Educational Studies*, in press, 54(1), 2006.

Carole J. Torgerson, M.Litt., *Senior Research Fellow*

Department of Educational Studies
University of York
YORK YO10 5DD

ABSTRACT: *The term 'publication bias' usually refers to the tendency for a greater proportion of statistically significant positive results of experiments to be published and, conversely, a greater proportion of statistically significant negative or null results not to be published. It is widely accepted in the fields of healthcare and psychological research to be a major threat to the validity of systematic reviews and meta-analyses. Some methodological work has previously been undertaken, by the author and others, in the field of educational research to investigate the extent of the problem. This paper describes the problem of publication bias with reference to its history in a number of fields, with special reference to the area of educational research. Informal methods for detecting publication bias in systematic reviews and meta-analyses of controlled trials are outlined and retrospective and prospective methods for dealing with the problem are suggested.*

1. INTRODUCTION

Since the late 1990s, 'evidence-based' policymaking for education in the UK has become increasingly widely talked about and, to some extent, developed. A key element in this development is a renewed interest in systematic review methodology and methods, as epitomised by the establishing in 2000 of two initiatives to increase the number of systematic reviews undertaken in the field: the Campbell Collaboration Social, Psychological, Educational and Criminological Trials Register (C2-SPECTR) in the USA, and the Evidence for Policy and Practice Information and Co-ordinating Centre at the Institute of Education, University of London in the UK.

Systematic reviewing presents a transparent and replicable approach to locating, identifying and synthesizing all the research literature in any given field. Systematic reviews aim either to exhaustively search for a population of studies or to sample representatively from it (Smith, 1980; Torgerson, 2003). Two main potential threats to the validity of systematic reviews are reviewer selection bias and publication bias. Reviewer selection effects occur when the criteria for study inclusion are developed in such a way as to 'select' into the review a biased sample of published studies. Because systematic reviews use transparent and replicable inclusion and exclusion criteria they are less likely to be affected by biased or selective reporting of the research literature than traditional narrative reviews. Despite this, however, the results of systematic reviews can be biased if there is a significant problem with publication bias. Researchers have long suggested that the published studies in the social sciences represent a biased sample of all the studies that are carried out (Rosenthal, 1979; Smith, 1980).

It is the aim of this paper to describe the problem of publication bias with reference to controlled trials and suggest methods of identifying and dealing with the problem.

2. WHAT IS PUBLICATION BIAS?

Publication bias is one of a range of reporting biases (including language bias and citation bias) that can affect the results of systematic reviews and meta-analyses of trials (true or quasi-experiments), and has been widely reported in the methodological literature (Sterne *et al*, 2000). Also known as positive outcome bias or the file-drawer effect, publication bias refers to the tendency for a greater proportion of statistically significant positive results of experiments to be published and, conversely, a greater proportion of statistically significant negative or null results not to be published (Greenwald, 1975; Rosenthal, 1979; Hedges and Olkin, 1980; Light, 1983; Light and Pillemer, 1984; Dickersin *et al*, 1987; Iyengar and Greenhouse, 1988; Begg and Berlin, 1988; Dickersin, 1997; Dickersin, 2002; Fitz-Gibbon, 2004). It also manifests as the tendency for published studies to have higher effect sizes than unpublished studies (Smith, 1980; Kulik and Kulik, 1989; Durlak and Lipsey, 1991), and for published studies to have larger sample sizes. The smaller the study, the larger the intervention effect will be necessary to demonstrate a statistically significant effect (Lipsey and Wilson, 2001). Therefore publication bias, if present in a review, will be partly a function of sample size (Dear and Begg, 1992), and a meta-analysis

containing a large number of small studies will have an increased risk of publication bias (Begg and Berlin, 1988). Publication bias could be due either to researchers tending not to submit their non-significant results or to journal editors tending not to accept them for publication (Lipsey *et al*, 1985; Wilson and Lipsey, 2001; Begg and Berlin, 1988).

If publication bias exists in a field, researchers searching for potentially relevant studies to include in systematic reviews will find studies with significant positive results easier to retrieve than studies with significant negative results (Wilson and Lipsey, 2001). If positive results are more likely to be published, this will bias the review towards a positive result because published studies are likely to be 'overrepresented' in systematic reviews (Iyengar and Greenhouse, 1988; Lipsey and Wilson, 1993; Wilson and Lipsey, 2001; Smith, 1980). Publication bias is, therefore, a potentially major threat to the validity of systematic reviews. Note, however, publication bias also affects non-systematic reviews in addition to their, usually unacknowledged, identification bias.

3. HISTORY OF PUBLICATION BIAS

Many researchers have demonstrated that the problem of publication bias is widespread. There is a consensus that the problem exists and that it is serious (Begg and Berlin, 1988). Selective publication was identified as being a problem in meta-analyses of experimental studies in educational research over 40 years ago (Sterling, 1959; Smart, 1964).

I have undertaken an overview of the effects of publication bias in the literature. I have included some key articles from healthcare research and other areas, such as psychological research, but where possible I have selected examples from the educational literature.

The earliest three studies that I identified were by Smart, Cohen and Sterling. They demonstrated that the majority of published studies in the field of psychology had statistically significant findings (Sterling, 1959; Cohen, 1962; Smart, 1964).

In his early cross-sectional study Sterling (1959) demonstrated that, in four major psychological journals published in 1955 and 1956, there was a greater probability of the results of experiments being published if the relevant test of significance rejected the null hypothesis than if the test failed to reject the hypothesis. In order to demonstrate that research that yielded non-significant results was routinely not being published, Sterling searched all the issues of four journals for the period January to December of either 1955 or 1956. Out of a total of 294 articles, 286 rejected the null hypothesis (at the 0.05 level of significance) and only eight failed to reject the null hypothesis at this level of significance. Sterling concluded that, because research yielding non-significant results was not being published, such research could be repeated until eventually, by chance, a significant result would occur (a Type 1 error) and would consequently be published leading to erroneous conclusions about the effectiveness of the intervention.

A few years after Sterling, Cohen surveyed the *Journal of Abnormal and Social Psychology* for the two years 1960 and 1961. He analysed the 70 articles that involved major statistical tests (Cohen, 1962) for their power to detect small, medium and large effects, using 2-tailed tests (at the 0.05 level of significance). He found that the mean power values (i.e. the probability of rejecting false null hypotheses) over the 70 empirical studies were 0.18 (one chance in 5 or 6) for small effects; 0.48 (slightly less than a 50-50 chance) for medium effects; and 0.83 (approximately a 5 out of 6 chance) for large effects. Because virtually all of the trials in Cohen's review had found, as statistically significant, small to medium effect sizes, unpublished studies with non-significant findings must be missing from the review. Cohen concluded that the power of these studies was far too small (unless the effect sought was large) and had probably led to the failure to reject false null hypotheses. As published research under-represents the research undertaken in a field, it is likely that significant numbers of studies having non-statistically significant findings had not been published in this journal for the two years surveyed.

In another cross-sectional study which replicated Sterling's work, Smart (1964) demonstrated that unpublished studies (conference papers) contained a higher proportion of negative results than did studies published in psychological journals; and that PhDs in psychology that reported negative results were less likely to be published than were those with positive results. He coded and compared: all of the psychological experiments published in four journals in 1961 and 1962; a non-random sample of 100 PhD theses from 1962; and a random sample of 100 unpublished papers presented at the American Psychological Association in 1962. He concluded that the neglect of negative studies was due to non-submission by authors or to greater critical examination of experiments containing negative results by journal editors and peer reviewers.

In order to estimate bias against the null hypothesis, Greenwald (1975) surveyed the authors and reviewers of all the manuscripts processed by him as associate editor of the *Journal of Personality and Social Psychology* during a three-month period in 1973. He asked these authors and reviewers about the relative probability of submitting studies for publication that either rejected or accepted the null hypothesis. The results indicated a strong bias against accepting the null hypothesis, illustrated by the 0.49 probability of submitting a rejection of the null hypothesis for publication compared with the low probability of 0.06 for submitting an acceptance of the null hypothesis for publication. Greenwald confirmed these findings by examining every article published in the *Journal of Personality and Social Psychology* in the year 1972 to determine what proportion accepted the null hypothesis. He found that out of a total of 100 articles only 24 reported acceptance of the null hypothesis.

In 1980, Smith examined a sub-sample of 12 meta-analyses in the fields of educational, social and psychological research and found that the findings from journals were, on average, one-third of a standard deviation more skewed towards the rejection of the null hypothesis than findings reported in theses or dissertations: that is a mean effect size of 0.64 in the published literature compared with a mean effect size of 0.48 in the unpublished literature (Smith, 1980).

These early findings have been more recently confirmed in the fields of healthcare research (Dickersin, 2002) and psychological, educational and behavioural treatment research (Lipsey and Wilson, 1993).

In their tertiary review of meta-analyses of psychological, educational and behavioural treatment research, Lipsey and Wilson (1993) found a 'strong skew' towards positive effects. They included 302 meta-analyses in their review. Only 6 of these reported negative effect sizes and relatively few reported effect sizes around zero - 85% of all the effect sizes were greater than 0.2. Lipsey and Wilson tried to identify reasons for this positive skew in their data and concluded that there were a number of possible factors leading to bias: selection bias and publication bias. They looked to see if evidence for the latter could explain the strongly positive effects by examining the differences in effect sizes between published and unpublished studies. They analysed a subset of 92 meta-analyses that reported separate effect sizes for published and unpublished studies, and found that the published studies had mean effect sizes 0.14 standard deviations larger than the mean effect sizes of the unpublished studies. These data support the view that studies with larger effect sizes are more likely to be published because, all things being equal, they will be more likely to be statistically significant.

In 1995, Sterling *et al* replicated Sterling's earlier study (1959) of the percentage of published articles in four major psychology journals that rejected the null hypothesis. This time Sterling and colleagues (1995) looked at eight psychology journals and three medical journals for either 1986 or 1987. They found that publication patterns in 1986/7 were still consistent with publication bias and that there had been little change since the original study. In the eight psychology journals, 95.56 % of articles using tests of significance rejected the null hypothesis (compared with 97.28% in 1958). The authors concluded that the practice of psychology journals preferring positive to negative results had not changed over the thirty-year period between 1956 and 1986. In the paper the authors cite a letter from an editor of a major environmental/ toxicological journal explaining why a manuscript had been rejected:

> Unfortunately, we are not able to publish this manuscript. The manuscript is very well written and the study was well documented. Unfortunately, the negative results translate into a minimal contribution to the field. We encourage you to continue your work in this area and we will be glad to consider additional manuscripts that you may prepare in the future (cited in Sterling *et al*, 1995, p.109).

Kulik and Kulik examined four of their own meta-analyses in the field of educational research for evidence of publication bias (Kulik and Kulik, 1989). The meta-analyses were undertaken in the areas of computer-based instruction at elementary and secondary level; computer-based instruction at post-secondary level; ability grouping; and mastery learning systems. Kulik and Kulik compared the mean effect sizes in all four meta-analyses for unpublished reports, unpublished dissertations and published journal articles. They found that in two out of the four meta-analyses (in computer-based instruction) the mean effect sizes for published journal articles were higher than those for both unpublished reports and dissertations, but for the other two meta-analyses the mean effect sizes for unpublished reports were higher. However in all four meta-analyses effect sizes were higher in journal articles than in dissertations. Kulik and Kulik urged caution in the interpretation of their results, claiming that the explanation for the relationship of the difference in effect sizes in journal articles and dissertations was 'controversial' (p.272), and not necessarily attributable to publication bias, but more likely to be attributable to the relative inexperience of dissertation writers.

More recent methodological studies of publication bias have been published in the field of healthcare research (e.g. Egger *et al*, 2003). In healthcare research there is a large methodological effort into the field of publication bias, particularly in the area of reviews of randomized trials. Indeed, the issue has gained such prominence that recently major medical journals have announced they will not publish randomized trials where the protocols have not been registered in advance in a publicly accessible database. This step should, in the long run, prevent small positive trials being published whilst small negative ones are ignored.

There are examples of meta-analyses in the sphere of educational research where it seems probable that publication bias has affected the magnitude of the pooled effect size. For example, Torgerson *et al* (2003) in a systematic review in the area of adult literacy and numeracy research noted that the field was probably susceptible to publication bias and concluded that small, negative studies evaluating interventions in adult basic education had probably not been published or were probably not in the public domain. Similarly, in their meta-analysis of experimental research into the effectiveness of second-language instruction, Norris and Ortega (2000) discussed the issue of publication bias at length. However, they decided not to include unpublished studies, and cautioned the reader that it was likely that this would lead to 'serious' publication bias (p. 432). Subsequently Truscott (2004) argued that a number of factors had affected the strong positive effect of the review (d = 0.96), including publication bias, and concluded that Norris and Ortega's results were 'substantially inflated' (p.22).

In summary, since 1959 various methodological and empirical researchers have found significant evidence for a file drawer effect within education and the social sciences. To avoid this bias having a detrimental effect on systematic reviews it is absolutely essential that, first, the problem is recognised and second, steps are taken to ameliorate this source of bias.

4. INCLUDING UNPUBLISHED DATA

Because publication bias has a long and ignoble history, researchers have sought methods both of identifying the problem and minimising its effects.

In their 'practitioner's guide to meta-analysis', Durlak and Lipsey (1991) outline the six major steps involved in conducting an effective meta-analysis and emphasise the procedures critical to the validity of its conclusions. They criticise a common practice in meta-analysis, that of only including published studies on the basis that these will represent the most high quality research in a given field, and suggest that quality criteria should be pre-specified and applied equally to published and unpublished studies (Lipsey and Wilson, 1993).

Researchers can minimise the problem of publication bias by: extensive and exhaustive searching; and by including studies that are unpublished but in the public domain, by searching the electronic databases that contain unpublished studies: for example, Dissertation Abstracts International; the System for Information on Grey Literature in Europe (SIGLE); and Education Resources Information Center (ERIC).

A justification often given for excluding unpublished studies in a systematic review, even if identified, is as a quality check. For example, in their meta-analysis of studies evaluating systematic phonics instruction versus non-systematic or no phonics instruction, Ehri and colleagues (2001) sought only studies from peer-reviewed journals. The authors justified their decision on the basis that unpublished studies were more likely to be of a lower quality than published studies.

Even if we include all the grey literature there will still be studies that we simply cannot detect, either because they have not been picked up by even the most sensitive search, or because they do not appear in the databases. If these 'missing' studies have similar characteristics to the identified studies the only issue that their non-inclusion raises is the risk of increasing a Type II error: that is wrongly concluding there is no statistically significant effect, when in truth there is. But if the missing studies are systematically different from included trials then bias can result. It is important to consider methods of detecting such bias and remedying the situation.

5. DETECTING PUBLICATION BIAS

Once the review has been completed, researchers should attempt to detect publication bias retrospectively and, if found to be present in a review, attempt to correct for it. If publication bias is found to be present in a systematic review, researchers can attempt a sensitivity analysis to assess the potential impact of missing studies on their results.

Methods for detecting and assessing publication bias, of the kind described in this section, have a reasonably long history. Indications of the existence of publication bias can be detected using a graphical or statistical method. Funnel plots are a graphical method, first used in educational research (Light and Pillemer, 1984). The effect sizes are plotted against the sample sizes (or standard errors) on a graph. The fail-safe *n* test (Rosenthal, 1978; Rosenthal, 1979; Dear and Begg, 1992) is a statistical method where the number of zero effect studies required to reduce the result to non-significance ($p > 0.05$) or to reduce a large effect size to a small effect size is calculated.

Funnel plot

The simplest and most common method used to detect publication bias is through the use of a funnel plot. In a funnel plot the point estimate from each study is plotted against some measure of the precision of the study (usually the standard error or sample size). Those studies with the highest precision will appear high up on the y –axis. A plot with little or no evidence of a publication bias should look like an inverted funnel. The largest study will be at the apex of the funnel with the smaller, less precise studies fanning out in equal measure on both sides of the large study or studies. If there is a publication bias present we will observe one side of the funnel to be missing (usually the left hand side indicating negative or null trials missing). Sometimes publication bias might be indicated by a hollowing out of the centre of the funnel plot around the area of no effect. This occurs when statistically significant positive and negative studies are published but those without significant results are not. The results of the funnel plot should then be taken into account when interpreting the conclusions of a systematic review.

For example, a secondary analysis of only the randomized trials (i.e. excluding the controlled trials) included in the National Reading Panel's review of systematic versus non-systematic phonics instruction (Ehri *et al*, 2001) noted that there was evidence for possible publication bias, as demonstrated by using a funnel plot (Torgerson, 2003). In this case, small negative studies appear to be missing. This bias may have led to an over-estimate of the benefits of systematic phonics instruction, although this interpretation should be treated with caution because there were only thirteen randomized trials in total in the review (see below) and asymmetry in a funnel plot is *suggestive* of publication bias.

Limitations of the funnel plot

Asymmetry in a funnel plot may be due to factors other than publication bias. There may be substantive or methodological heterogeneity between the studies. An asymmetrical funnel plot may occur when small, methodologically weak trials produce biased estimates of effect and consequently appear as 'positive' trials when they should be null or negative studies. Sterne *et al* (2000) have outlined four possible reasons for asymmetry in funnel plots: one of these is reporting (publication) bias; the others are true heterogeneity, data irregularities (poor methodological design) and chance. The possibility of chance accounting for an asymmetrical funnel plot will increase with a declining sample size of trials. Therefore, funnel plots with fewer than 20 trials should be interpreted with caution as an asymmetrical funnel plot may have occurred simply because no trial with an extreme result has yet been conducted.

Comparisons between published and unpublished studies

As well as using funnel plots we can also look at the effect sizes between unpublished and published data. For example, in their meta-analysis of adult literacy and numeracy trials, Torgerson *et al* (2003) noted much larger effect sizes among published data ($d = 0.49$, $p=.003$) compared with unpublished studies ($d = 0.26$, $p=.13$). Similarly, Lipsey and Wilson (1993) also found that unpublished studies had an average effect size somewhat smaller than the average effect size of the published data.

Fail-safe n test

If there is a suspicion of publication bias then the results of the systematic review can be subjected to the fail-safe *n* test. This is the test that determines the number of studies not retrieved averaging an effect size of zero that would need to exist in order to reduce the summary effect to a non-significant level or to bring the overall probability of a Type 1 error to a stated level of significance, such as $p=0.05$ (Rosenthal, 1978; Rosenthal, 1979). Rosenthal (1979) indicated that the findings of a meta-analysis are probably robust if the fail-safe *n* is not more than five times the number of reviewed studies plus ten. In their meta-analysis of studies evaluating systematic phonics instruction versus non-systematic or no phonics instruction Ehri *et al* (2001, p.431) calculated that, for the 43 comparisons they found (in RCTs and CTs) with effect sizes of $d = 0.20$ or greater to be 'statistical exceptions', the existence of 860 comparisons in the unpublished literature with effect sizes below $d = 0.20$ would be required, and they considered this possibility 'unlikely'. However, in his meta-analysis of 11 studies evaluating the effects of reading to young children in schools, Blok (1999) found pooled effect sizes of 0.41 for reading and 0.63 for oral language development. He calculated that the fail safe *n* for the oral language outcome was 22, i.e. 22 zero effect unpublished studies would be required to reduce this effect size from 0.63 to 0.2. Because this

number is not more than five times the number of reviewed studies plus 10 (i.e. 65) it is conceivable that there are sufficient unpublished trials to overturn the result.

Limitations of the fail-safe n test
This method is based on the assumption that the unpublished studies are a random sample of all the studies that were undertaken (Iyengar and Greenhouse, 1988). It combines the results of the selected studies as if they were an unselected sample and then retrospectively assesses the potential effects of publication bias. This assumption is, of course, unlikely ever to be strictly true. The method also over-emphasises the importance of the convention of using p=0.05 to test for statistical significance.

6. CONCLUSIONS

The problem of publication bias was first recognized in the field of psychology, although many authors have raised the issue of publication bias over the last 40 years in other fields, in particular in healthcare research. Much methodological work has recently been undertaken in this field (see, for example, Egger *et al*, 2003) and in some areas of the psychological and the social sciences (Sterling *et al* 1995, Lipsey and Wilson, 1993) to demonstrate the existence of the problem, and to illustrate ways of correcting for it retrospectively.

Sutton *et al* (2000) and Egger *et al* (2003) have demonstrated that, in the field of healthcare research, many meta-analyses do not consider the effect of publication bias on their results. Sutton *et al* analysed 48 systematic reviews in the Cochrane Database of Systematic Reviews. 23 meta-analyses were estimated to have some degree of publication bias (with random effects model). The authors estimated that about half of meta-analyses may be subject to some level of publication bias and about a fifth have a strong indication of missing trials. This analysis concluded that publication bias was common within the sample of meta-analyses, but that in most cases the bias did not affect the conclusions. The authors deduced that around 5-10% of meta-analyses could have been interpreted incorrectly because of publication bias.

The issue of publication bias is an important threat to evidence informed research and policy-making. For example, in a systematic review and meta-analysis of basic adult literacy interventions, an overall benefit of intervention was observed. However, the funnel plot suggested that there was evidence for possible publication bias so the authors concluded that the results should be treated with caution and ideally be confirmed in a large RCT (Torgerson *et al*, 2003).

Publication bias is an important threat to the validity of systematic reviews. Researchers undertaking reviews need to be aware of the problem and investigate the possibility of publication bias in their review. Readers of systematic reviews should always be aware that a review containing lots of small positive trials is particularly threatened by publication bias. Many of these small trials may have false positive results, and it is likely that small trials containing negative results (whether false or true) have been undertaken but have not been included in the review. There are a number of ways in which the problem can be limited prospectively. In order to attempt to prevent the problem, journal editors should encourage the submission of good quality, but negative or null studies. A recent, even more extreme, suggested development is to have journals dedicated to the publication of null results. Researchers have a responsibility to ensure the timely submission of their trials for publication, whatever their results. In the field of healthcare research it has been suggested that a way of limiting the consequences of publication bias is to set up trial registries in all areas of research and critically assess the process of peer review (Begg and Berlin, 1988). The process of setting up trial registries is well underway for healthcare trials. Such a system would reduce the problem of publication bias in research in education and the social sciences.

7. ACKNOWLEDGEMENTS

I thank Greg Brooks (University of Sheffield), Stephen Gorard (University of York), the two anonymous referees and the editor of the journal for their helpful comments on an earlier draft of this paper.

8. REFERENCES

BEGG, C.B. and BERLIN (1988), J.A. Publication bias: A problem in interpreting medical data, *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 151 (3): 419-463.
COHEN, J. (1962), The statistical power of abnormal-social psychology research: A review, *Journal of Abnormal and Social Psychology*, 65(3): 145-153.

DEAR, K.B.G. and BEGG, C.B. (1992) An approach for assessing publication bias prior to performing a meta-analysis, *Statistical Science*, 7(2): 237-245.
DICKERSIN, K., CHAN, S., CHALMERS, T.C., SACKS H.S. and SMITH, H. (1987) Publication bias and clinical trials, *Controlled Clinical Trials*, 8: 343-353.
DICKERSIN, K. (1997) How important is publication bias? A synthesis of available data, *AIDS Education and Prevention*, 9 (Supplement A): 15-21.
DICKERSIN, K. (2002) *Reducing reporting biases*, in Chalmers, I., Milne, I., Trohler, U. (eds.) The James Lind Library (www.jameslindlibrary.org).
DURLAK, J.A. and LIPSEY, M.W. (1991) A practitioner's guide to meta-analysis, *American Journal of Community Psychology*, 19(3): 291-332.
EGGER, M., JUNI, P., BARTLETT, C., HOLENSTEIN, F. and STERNE, J. (2003) How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? *Empirical Study Health Technology Assessment*, 7(1): 1-76.
EHRI, L.C, NUNES, S.R., STAHL, S.A. and WILLOWS, D.M., (2001) Systematic phonics instruction helps students learn to read: evidence from the national reading panel's meta-analysis, *Review of Educational Research*, 71: 393-447.
FITZ-GIBBON, C. (2004) Editorial: The need for randomized trials in social research, *Journal of Royal Statistical Society*, 167(1): 1-4.
GREENWALD, A.G. (1975) Consequences of prejudice against the null hypothesis, *Psychological Bulletin*, 82(1): 1-20.
HEDGES, L.V. and OLKIN, I. (1980) Vote counting methods in research synthesis, *Psychological Bulletin*, 88(2): 359-369.
IYENGAR, S. and GREENHOUSE, J.B. (1988) Selection models and the file drawer problem, *Statistical Science*, 3(1): 109-117.
KULIK, J.A. and KULIK, C-L.C. (1989) Meta-analysis in education, *International Journal of Educational Research*, 13: 221-340.
LIGHT, R.J. and PILLEMER, D.B. (1984) *Summing up: the science of reviewing research*, Cambridge, MA: Harvard University Press.
LIPSEY, M.W., CROSSE, S., DUNKLE, J., POLLARD, J. and STOBART, G. (1985) Evaluation: The state of the art and the sorry state of the science, *New Directions for Program Evaluation*, 27.
LIPSEY, M.W. and WILSON, D.B. (1993) The efficacy of psychological, educational and behavioral treatment: confirmation from meta-analysis, *American Psychologist*, 12: 1181-1209
NORRIS, J.M. and ORTEGA, L. (2000) Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis, Language Learning, 50(3):417-528.
ROSENTHAL, R. (1978) Combining results of independent studies, *Psychological Bulletin*, 85: 185-193.
ROSENTHAL, R. (1979) The 'file drawer problem' and tolerance for null results, *Psychological Bulletin*, 86(3): 638-641.
SMART, R.G. (1964) The importance of negative results in psychological research, *Canadian Psychologist*, 5: 225-32.
SMITH, M.L. (1980) Publication bias and meta-analysis, *Evaluation in Education*, 4: 22-24.
STERLING, T.D. (1959) Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa, *Journal of American Statistical Association*, 54: 30-34.
STERLING, T.D., ROSENBAUM, W.L. and WEINKAM, J.J. (1995) Publication decisions revisited: the effect of the outcome of statistical tests on the decision to publish and vice versa, *The American Statistician*, 49(1): 108-112.
STERNE, A.C., GAVAGHAN, D. and EGGER, M. (2000) Publication and related bias in meta-analysis, *Journal of Clinical Epidemiology*, 53(11): 1119-1129.
SUTTON, A.J., DUVAL, S.J., TWEEDIE, R.L., ABRAMS, K.R.A. and JONES, D.R. (2000) Empirical assessment of effect of publication bias on meta-analyses, *British Medical Journal*, 320: 1574-1577.
TORGERSON, C.J. (2003) *Systematic reviews*. London: Continuum Books.
TORGERSON, C.J., PORTHOUSE, J., BROOKS, G. (2003) A systematic review and meta-analysis of randomised controlled trials evaluating interventions in adult literacy and numeracy, *Journal of Research in Reading*, 26(2): 234-255.
TRUSCOTT, J. (2004) The effectiveness of grammar instruction: analysis of a meta-analysis, *English Teaching and Learning*, 28(3):17-29.
WILSON, D.B. and LIPSEY, M.W. (2001) The role of method in treatment effectiveness research: evidence from meta-analysis, *Psychological Methods*, 6(4): 413-429.

Appendix B

Carole J Torgerson, David J Torgerson, Yvonne F Birks, Jill Porthouse A Comparison of Randomised Controlled Trials in Health and Education, *British Educational Research Journal*, in press, 2005.

Department of Educational Studies
University of York
YORK YO10 5DD
Carole Torgerson, Senior Research Fellow

Department of Health Sciences
University of York
YORK YO10 5DD
David Torgerson, Professor
Yvonne Birks, Research Fellow
Jill Porthouse, Research Fellow

Correspondence to: Carole Torgerson
Email: cjt3@york.ac.uk

Abstract

Background: Healthcare and educational trials face similar methodological challenges. Methodological reviews of healthcare trials have shown that a significant proportion has methodological flaws. Whether or not educational trials have a similar proportion of poor quality trials is unknown. We undertook a methodological comparison between healthcare and educational trials published since 1990.

Aims: (1) To assess whether the quality of trial reports in education and healthcare are similar. (2) To assess whether the quality of trial reporting is improving.

Methods, Results: The characteristics of a sample of trials, published since 1990, were taken from health and educational journals. Trials were assessed using the following quality criteria: rationale for sample size; concealment of allocation; blinded follow-up; use of confidence intervals; adequate sample size. We identified 96 placebo drug trials and 54 non-drug trials published in major general journals. We compared these with 54 trials in specialist health journals and 84 trials in educational journals. No educational trial used concealed allocation or reported the rationale for sample size calculation and only one trial used confidence intervals. There was a trend for the reporting of healthcare trials to improve with time, whilst the reporting quality of educational trials declined.

Conclusion: Poor quality of trial reporting is more prevalent in educational journals than in healthcare journals.

Background

Over the last 20 years reviews have highlighted methodological weaknesses of many randomised controlled trials (RCTs) of healthcare interventions published in major medical journals (Pocock *et al*, 1987; Altman & Dore, 1990; Gore *et al*, 1992; Schulz *et al*, 1994; Moher *et al*, 1994; Schulz *et al*, 1995; Assman *et al*, 2000). Partly in response to these concerns many journals have adopted the CONSORT statement in an attempt to improve the reporting of trials (Begg *et al*, 1996).

Part of the need to develop the CONSORT statement arose from the widespread use of systematic review and meta-analysis methods in healthcare. Ideally in a meta-analysis trials using similar methods need to be identified for the purposes of 'pooling' the results. In order to identify a sample of homogeneous trials the trial methods need to be clearly and correctly described. As well as allowing the identification of similar trials CONSORT also allows meta-analysts to exclude studies that are so methodologically weak as to be potentially fatally flawed. Similarly, researchers in education are returning to systematic review methods and meta-analysis (Torgerson, 2003). Therefore, the issue of identifying high quality trials for potential inclusion in a synthesis is important in this field as well.

Whilst, arguably, trials are undertaken in healthcare more often than in any other specialism there are areas where the use of the controlled trial has a longer history. Indeed, agricultural trials pre-date healthcare trials by many years (Oakley. 2000). Similarly, in educational research the use of trials has

been widely used for a long period of time (Oakley, 2000). It is unknown whether or not similar problems of trial quality affect educational research as well as healthcare trials.

The quality of educational trials is likely to become of increasing importance if, and when, educational policy makers start to demand evidence for policy changes based on randomised data. Indeed, there are increasing calls from some educational researchers to undertake high quality randomised trials (Boruch, 1994; Torgerson & Torgerson, 2001; Torgerson *et al*, 2003). Furthermore, when trials are put into a meta-analysis it is important that their characteristics and methods are clearly described (Torgerson, 2003). It is therefore important that educational trials should be of high quality. In this paper we compare the quality of a sample of trials undertaken in both healthcare and education since 1990.

Methods

We wanted not only to make an absolute assessment of the quality of recent educational trials but also to undertake a relative quality assessment against 'control' groups. Therefore, we decided to choose trials in healthcare as a comparator group. We felt it was important to have a 'control' group in order to see whether or not trials facing similar methodological challenges, but in a different setting, could overcome them and improve the quality of the design and conduct of experiments. Furthermore, many healthcare trials have an educational component, albeit related to health, and therefore we could make appropriate comparisons. Nevertheless, many healthcare trials, such as placebo controlled drug evaluations or surgical trials, are not directly comparable with studies in educational research we decided to choose non-surgical and non-pharmaceutical trials for our main comparison. Such trials have some important similarities to educational trials. First, they cannot generally be double blinded. Both the participant and the researcher are usually aware of the allocated treatment. Second, outcome is often measured using non-physiological measures of outcome, such as quality of life or changes or indeed, educational test scores (albeit with a health focus). To see if our results would be influenced by the choice of open trials we also identified a sample of placebo controlled drug trials from major medical journals to act as a reference group. We decided also to only include trials published in the English language.

We hand searched the *Lancet*, the *British Medical Journal*, *Journal of the American Medical Association* and the *Archives of Internal Medicine* for open pragmatic trials (12 references each). Open trials are those that do not use a placebo or sham treatment and in which the participants are aware of their allocated interventions. The choice of journals for these 'open' trials was driven by the knowledge these high impact healthcare journals have a history of publishing 'open' non-drug trials. For placebo controlled trials we searched the *British Medical Journal*, *Lancet*, *New England Journal of Medicine* and the *Annals of Internal Medicine* (i.e. 96 papers). In order to assess the potential for selection bias in our choice of healthcare journals we also identified a sample of open trials concerned with healthcare educational interventions) among children (e.g. asthma management education) from either the Cochrane database, which aims to reference all RCTs in healthcare, or by handsearching accessible journals. To avoid a single journal dominating our results we sought only 2 placebo trials and 1 open trial from each year in a given journal.

The number of published educational trials is substantially fewer than in the field of healthcare research. Therefore, in order to obtain an adequate sample of trials we could not restrict ourselves to a few 'high profile' journals. For relevant educational trials we used studies identified by completed systematic reviews by some of the authors (Torgerson & Elbourne, 2002; Torgerson *et al*, 2002; Torgerson *et al*, 2003; Torgerson & Zhu, 2003). We supplemented trials identified from these systematic reviews through hand searches of key educational journals. We included 20 references from the following hand searched journals: *The British Journal of Educational Psychology* (10 references); *Educational Research* (3 references); *Journal of Research in Science Teaching* (7 references). As with the health trials we only took one article from the same journal in any year. Initially, we looked at C2SPECTR, which is an emerging database of controlled trials in criminal justice, social policy and education as a source of trials. However, at the time of our search the database did not include most recent trials and therefore, we opted for a combination of hand searches and the results from ongoing electronic searches of other databases.

We chose the following five markers of trial analysis, methodology and reporting quality: Was the method of randomisation concealed? Was there a justification for the sample size? Was there an adequate sample size? Was there blinded follow up? Did the authors use confidence intervals?

Sample size

The choice of a sample size of any trial is very often an arbitrary process. One of the key issues in the use of a sample size calculation is the potential to minimise the chances of us making a 'Type II' error. A Type II error occurs when there is a true difference between the randomised groups; however, this difference is either not statistically significant or is simply missed. The probability of a Type II error

occurring declines with an increasing sample size: larger trials, all other things being equal, are less likely to experience a Type II error than smaller studies. Typically we want to have a sample size that will have a low probability of missing an educationally (or in healthcare, clinically) important difference. One of the key problems with respect to sample size estimation is the definition of what difference is important. For the purposes of this paper we propose that a half a standard deviation difference (i.e., 0.5 effect size) is the largest difference that we should power our trials to detect. The justification for this is that an overview of quasi-experimental research in the educational and psychological literature noted that most interventions had an effect size of 0.5 or smaller (Lipsey and Wilson, 1993). Few commonly evaluated interventions gave larger educational effects. Similarly, if the outcome is a binary variable it is unlikely that any intervention will lead to greater than a doubling or halving of effect. The other issues that affect sample size are: power and significance. Traditionally in health research and most of the social sciences a significance value of 5% is generally recognised as being statistically significant. The notion that the 5% value should be statistically significant was introduced by Fisher more than 60 years ago (Sterne and Davey-Smith, 2001). This value is purely arbitrary as is the related phenomenon of using 95% confidence intervals. Indeed, it has been argued recently, in the context of healthcare, that 90% confidence intervals should be used as well as 5% p values (Sterne and Davey-Smith, 2001), although any choice of a 'threshold' value will be arbitrary. However, rather than discussing this issue in detail we propose to use the 'traditional' notions of power and statistical significance for setting a sample size: that is at a 5% significance level with 80% power. Therefore, for continuous outcomes we assumed that a difference of half a standard deviation would be required (i.e. a sample size of 126). The actual sample size of the trial was then subtracted from the minimum sample size, which was calculated using commercially available software. For cluster trials to account for the intra cluster correlation co-efficient we assumed that the sample size ought to be 50% larger than a similar individually randomised study.

Results

We identified 96 placebo controlled trials and 54 open trials in the major general journals. We also identified 84 education trials published in 43 different journals and 35 open non-drug trials.

Table 1 presents the characteristics of the trials we identified. Healthcare trials were more likely than educational trials to report all of the quality measures. Nevertheless, healthcare trials not published in major journals did not have as high a prevalence of markers of quality as studies published in high profile general medical journals.

To assess whether there was any association over time between the reporting quality of a trial and the time it was published we correlated the mean number of reported items by year of publication. For placebo controlled drug trials there was a positive (0.39) and statistically significant correlation ($p<0.01$). A similar result was found for the 'open' trials published in the general major journals (0.39, $p<0.01$); however, for the sample of trials taken from 'other' journals there was a slight, not statistically significant decline in quality (-0.02, $p=0.91$). For trials published in the educational literature there appeared to be a decline in quality (-0.21, $p=0.06$) of trials published over time.

In table 2 we explore the relationship between year of publication and the presence or absence of our quality criteria. The table shows that for healthcare trials there is a tendency for all markers of quality to improve with time, the exception being underpowered trials in specialist journals where there was a tendency for a decline in appropriately powered studies. In contrast, however, educational trials showed a significant decline in the numbers of trials reporting blinded follow-up and a tendency for trials to decline in statistical power. For two of the quality criteria (concealed allocation and sample size justification) it was not possible to look at changes over time as no study reported these. Only one educational trial used confidence intervals in the reporting of the results.

Discussion

Trials published in the health literature appear to be of higher quality than studies taken from education. The stronger performance of health trials in this comparison is partly explained by the fact that the studies we chose were published in the higher quality journals. Nevertheless, even when we sought trials from specialist journals the quality differential, although weakened, still persisted.

Trials published in the major healthcare journals also seemed to be improving over time, which was not the case with educational trials. These actually appeared to be declining in quality. However, we did not notice an improvement among the reporting quality of healthcare trials that were published in specialist healthcare journals. This may be because fewer specialist journals are CONSORT members than general journals. In placebo controlled drug trials all our markers of quality appear to be improving with time with the exception of blinded follow-up. This latter finding is probably due to fact that it is relatively

easy to blind follow up due to the nature of the trial, indeed, this is one of justifications for using placebos.

It is interesting to note among educational trials a significant reduction in the use of blinded follow-up with time. Blinding or masking assessors is important because this avoids reporting or ascertainment bias and has been the hallmark of good trial design, including non-health studies, for many years (Cook & Campbell, 1979).

Very few educational trials met our pre-specified sample size. Sample size is an important consideration when designing a study. Small trials can and will miss important effects. Whilst meta-analysis can and does go some way to addressing the problem of small sample sizes by pooling together similar trials this is an imperfect solution for a number of reasons. One important problem with small trials is that those, which by chance, produce a negative or null result are less likely to be published than those that are positive. This in turn will bias the results of a meta-analysis by either overestimating a positive effect or, worse, erroneously concluding an intervention has a positive effect when it does not. Larger trials, on the other hand, whatever their results are more likely to be published. They are also more likely to give an estimate that is closest to the 'correct' answer than smaller studies. The finding that 85% of educational trials did not have adequate power to show as statistically significant (albeit at the arbitrary 5% value) is a cause for concern. It may be better for educational researchers, rather than undertaking a series of small trials, to collaborate more closely and undertake a large, multi-centred study, which with similar inclusion/exclusion criteria will produce much more powerful results than a meta-analysis of a series of disparate small studies.

It is important to take into consideration that, apart from sample size and the use of confidence intervals, the other markers of trial quality are dependent on the authors' reporting. It is possible that in some trials, for both health and education, important aspects of trial quality such as blinded outcome assessment could have been undertaken but not reported by the authors. This may have led to some bias in our results. Because general medical journals, since 1996, insist that important markers of trial quality are reported in the published article this may have lea to an over-estimate of the quality of trials published in these journals compared with specialist health journals and educational journals, both of which do not, at present, adopt a policy of trial reporting. This problem of under-reporting of trial quality could be addressed by specialist health journals adopting the CONSORT statement and by editors of educational journals developing a similar checklist to quality assure the publication of educational trials. This would not only help those who review trials in education but also act as an incentive for future educational trials to be more rigorously designed and executed.

Table 1: Characteristics of identified 'open' health and educational trials

Characteristics	Placebo Drug Trials Major Journals (n = 96)	Pragmatic Health Trials Major Journals (n = 54)*	Education (n = 84)	Open Health 'Other' Journals (n = 34)
Cluster randomised	1 (1.0%)	11 (20.4%)	15 (17.9%)	21 (61.8%)
Dichotomous outcome	50 (50.0%)	26 (48.1%)	6 (7.1%)	7 (20.6%)
Rationale for Sample size	57 (59.4%)	21 (38.9%)	0	4 (11.8%)
Allocation Concealment	38 (39.6%)	6 (11.1%)	0	1 (2.9%)
Blinded Follow-up	51 (53.1%)	19 (35.2%)	12 (14.3%)	7 (20.6%)
Use of Confidence Intervals	65 (67.7%)	27 (50.0%)	1 (1.2%)	9 (26.5%)
Adequately Powered	53 (55.2%)	29 (53.7%0	13 (15.5%)	22 (64.7%)

*Includes two trials, one each from *New England Journal of Medicine* and *Annals of Internal Medicine,* identified by search of Cochrane database.
NB – A list of included trials is available from the authors on request.

Table 2: Increased odds of characteristic for per year

Characteristics	Placebo Drug Trials Major Journals (n = 96)	Pragmatic Health Trials Major Journals (n = 54)*	Education (n = 84)	Open Health 'Other' Journals (n = 34)
Rationale for Sample size	1.20 (1.05 to 1.36) p=0.006	1.24 (1.04 to 1.46) p=0.014	NC	1.04 (0.76 to 1.43) p=0.81
Allocation Concealment	1.13 (1.00 to 1.28) p=0.06	1.00 (0.80 to 1.26) p=0.98	NC	1.47 (0.47 to 4.62) p=0.51
Blinded Follow-up	1.04 (0.93 to 1.17) p=0.50	1.12 (0.95 to 1.30) p=0.18	0.76 (0.61 to 0.94) p=0.013	1.09 (0.84 to 1.42) p=0.50
Use of Confidence Intervals	1.20 (1.05 to 1.38) p=0.008	1.22 (1.40 to 1.44) p=0.016	1.89 (0.49 to 1.63) p=0.36	1.05 (0.84 to 1.32) p=0.66
Adequately Powered	1.20 (1.06 to 1.37) p=0.005	1.08 (0.93 to 1.25) p=0.33	0.95 (0.80 to 1.13) p=0.56	0.81 (0.64 to 1.03) p=0.08

NC = Not calculable.
NB An Odds Ratio of greater than 1.0 indicates an increased reporting of that quality characteristic.

References

Altman DG, Dore CJ. (1990) Randomisation and baseline comparisons in clinical trials. *Lancet* 335:149-53.

Assman SF, Pocock SJ, Enos LE, Kasten LE. (2000) Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 355:1064-69.

Begg CB, Cho MK, Eastwood S, *et al.* (1996) Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *Journal of American Medical Association* 276:637-39.

Boruch RF. (1994) The future of controlled randomized experiments: a briefing. *Evaluaton Practice* 15:265-74.

Cook TD, Campbell D. (1979) Quasi-Experimentation: Design and Analysis Issues for Field Settings.

Gore SM, Jones G, Thompson SG. (1992) The Lancet's statistical review process: areas for improvement by authors. *Lancet* 340:100-2.

Lipsey MW and Wilson DB (1993) The Efficacy of Psychological, Educational, and Behavioral Treatment. Confirmation From Meta-Analysis. *American Psychologist,* December 1993.

Moher D, Dulberg CS, Wells GA. (1994) Statistical power, sample size, and their reporting in randomized controlled trials. *Journal of American Medical Association* 272:122-24.

Oakley A. *Experiments in Knowing: Gender and Method in the Social Sciences.* Polity Press: London, 2000.

Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. *New England Journal of Medicine* 1987;317:426-32.

Schulz KF, Chalmers I, Grimes DA, Altman DG. (1994) Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. *Journal of American Medical Association* 272:125-28.

Schulz KF, Chalmers I, Hayes RJ, Altman DG. (1995) Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of American Medical Association* 273:408-12.

Sterne JAC, Davey-Smith G. Sifting the evidence – what's wrong with significance tests? (2001) *British Medical Journal* 322;226-31.

Torgerson CJ, Elbourne D. (2002) A systematic review and meta-analysis of the effectiveness of information and communication technology (ICT) on the teaching of spelling. Journal of Research in Reading. 25:129-43.

Torgerson CJ, Sowden A, King SE. (2002) Do Volunteers in Schools Help Children Learn to Read? A Systematic Review of Randomised Controlled Trials. *Educational Studies* 28: 433-44.

Torgerson CJ, Torgerson DJ. (2001) The need for randomised controlled trials in educational research. *British Journal of Educational Studies* 25:129-43.

Torgerson DJ, Torgerson CJ. (2003) The Design and Conduct of Randomised Controlled Trials in Education: lessons from healthcare. *Oxford Review of Education* 29:67-80.

Torgerson CJ, Porthouse J, Brooks G. (2003) A Systematic Review of Interventions to Promote Adult Literacy and Numeracy. *Journal of Research in Reading* **26**:234-55.

Torgerson CJ, Zhu D. (2003) A systematic review and meta-analysis of the effectiveness of ICT on literacy learning in English. (EPPI-Centre Review) In: *Research Evidence in Education Library.* Issue 2. London: EPPI- Centre, Social Science Research Unit, Institute of Education.

Torgerson C. (2003) *Systematic Reviews* Continuum Books: London.

Commentary Two:

Bias in systematic reviews: conclusions and recommendations

Contents

Recapitulation

This portfolio has outlined the rationale and methods of systematic review (Item 1). It has systematically examined the use of systematic reviews of randomized controlled trials to inform policy, practice and scholarship, in the field of literacy learning (Item 2). Finally, it has used a systematically assembled dataset of meta-analyses and their included RCTs to undertake methodological work in areas that pose potential threats to the validity of both trials and systematic reviews of trials (Item 3). The three main research questions of the portfolio were:

(1) Are systematic reviews of randomized controlled trials in literacy research of high quality?
(2) Is the quality of the randomized controlled trials in literacy research fit for purpose?
(3) Are indicators of the quality of randomized controlled trials associated with measures of effectiveness?

The portfolio has addressed the three research questions posed. Specifically, it has found that systematic reviews that include RCTs in literacy research are of generally good quality, that many RCTs in literacy research are, however, not fit for purpose, and finally that, despite this, there is little evidence to show an association between quality of RCTs and outcome.

There are some apparent contradictions in the conclusions to this portfolio. Applying the QUORUM criteria to systematic reviews in literacy research in Item 2 appeared to show that these are of good quality. Nevertheless, further in the portfolio (Item 3,

Chapter 4) replication of one of the reviews lead to a reduced confidence in the quality of this review. This may be because QUORUM is a set of guidelines that reflects *reporting* quality and not necessarily the quality of the review itself. Only replication of a review by an independent research group can achieve the 'gold-standard' of quality assurance.

The trials tended to be of poor quality, yet no statistically significant association was found between observable measures of quality (e.g., sample size, attrition rates) and effect size. This may be due to the relatively small sample of trials involved in the comparisons leading to possible Type II errors. Some aspects of trial quality, such as allocation methods, were unobservable variables because they were simply not reported by the trialists. For example, during the replication of the National Reading Panel phonics review it was found that, in one of the trials, not only was it impossible to tell from the published article whether or not the participants had been allocated to intervention or control groups by randomization, the lead author herself was unsure until she revisited her records. It may be that trialists have simply misreported aspects of their trial methodology, which will have masked any associations between quality and outcome.

In summary, therefore, the QUORUM statement showed that literacy systematic reviews were reported well, but replication of a single review casts doubt on the validity of at least one of the reviews. Trial quality was poor but this did not seem to be associated with effect size.

Conclusions

This portfolio is the first body of work which, to the best of my knowledge, systematically examines the quality of systematic reviews of randomized trials, and the quality of the included trials themselves, in the field of literacy learning.

The methodological work undertaken in Item 3, Chapters 3 and 4 is unique. To the best of my knowledge it has not been demonstrated previously that there is little evidence of publication bias in literacy research. Item 3, Chapter 2 reviewed the literature on publication bias and found no evidence of previous methodological work in the field of literacy learning.

The methodological work undertaken in Item 3, Chapters 6 and 7 makes an important contribution to the topic of trial quality in the field of educational research in general and of literacy learning in particular. I have demonstrated that one set of criteria adopted for critical appraisal of RCTs is insufficient and that, when modified criteria based upon the CONSORT statement are applied to trials in phonological awareness training, these show that many of those trials are methodologically weak. I have applied a subset of these CONSORT criteria to a sample of 84 trials sampled from the educational literature over a 12-year period. I have shown that these trials, like those in the field of phonological awareness instruction, are methodologically weak. I have also shown that there has been a decline in methodological rigour over the time period. However, I have demonstrated that the modified CONSORT criteria are adequate for the quality appraisal of the internal validity of randomized controlled trials in literacy learning. Trials in literacy learning are, in general, poorly reported. Most do not report

the key aspects of a study design and conduct. Therefore, it is difficult to quality appraise such trials. Because quality appraisal is so difficult, it is problematic to rely too heavily on current evidence to support different approaches to literacy learning.

Summary of conclusions from each item of the portfolio:

Item 1:

- In the UK and at the beginning of the 21st century there is an increasingly high profile of 'evidence-based' research in education to inform policy, practice and scholarship. A key element of this movement is a renewed focus in education on systematic reviews, to limit potential selection, publication and other biases in the synthesis of evidence.
- The most robust methods for probing effectiveness questions in educational research are randomized controlled trials and systematic reviews of randomized controlled trials.
- The most important aspects of trial design relate to internal validity. The quality of randomized controlled trials should, amongst other things, be assessed on their internal validity. The CONSORT statement (Altman, 1996) for the quality appraisal of RCTs is a useful tool for this purpose.
- Publication bias can threaten the validity of systematic reviews; therefore it is important that its presence is detected and discussed, as part of a detailed quality assessment of systematic reviews in any given field. The QUORUM statement (Moher *et al*, 1999; Shea *et al*, 2001) is a useful tool for the quality appraisal of systematic reviews.

Item 2: Quality of studies included in the tertiary review

- In terms of the first research question, most of the reviews in the tertiary review seem to be of high quality, according to their quality appraisal using the modified QUORUM statement. All the reviews clearly stated their research question, and their methods of searching and selecting included studies. Most studies described their data extraction and used some form of quality assessment of included studies.
- On the other hand, some reviews did have notable methodological weaknesses. Six of the 14 studies did not make an assessment of publication bias, which is potentially a major threat to the validity of any systematic review. In addition, six studies did not provide evidence for reviewer agreement when synthesising the data. There is, therefore, some room for improvement in the methodological quality of systematic reviews in literacy learning.
- Some reviews included both RCTs and CTs. When examining the effect sizes of RCTs and CTs separately there was no clear pattern as to whether the RCTs produced larger or smaller effect sizes than the CTs.

Item 2: Pedagogical findings of studies included in the tertiary review

- Because the reviews were judged to be of generally high quality according to the QUORUM guidelines, the pedagogical findings of the reviews were considered to be reliable.
- There was little evidence of benefit from the use of information and communication technology (ICT) in spelling or reading instruction.
- However, the use of word-processing for improving the quality of writing seems to be a promising intervention, particularly for weaker writers.

- There was some evidence to support the use of interventions involving the use of adults reading aloud to children in both home and school settings.
- There was sound evidence of benefit of structured approaches to phonemic awareness instruction and systematic phonics instruction in early literacy, particularly in populations of children at risk of reading failure.
- There was some evidence that whole language instruction in reading was not beneficial, compared with 'basal instruction' (initial teaching using reading schemes).
- The evidence for the use of volunteers in early reading development was equivocal: one review found a benefit of the use of one-to-one volunteers for helping 'struggling' readers; a second review, by contrast, found a small positive effect, but the difference between the groups was not statistically significant.
- The use of writing instruction in narrative and expository text was found to be beneficial in one review.
- There was a substantial benefit of meta-cognitive instruction on reading comprehension. Awareness of textual consistency and the use of self-questioning for monitoring and regulating were found to be the most beneficial strategies (based on a single review, however).
- A single review of peer-tutoring found a positive effect on reading outcomes.

Item 3: Publication bias

- Contrary to other reviews, mainly in healthcare research and psychology, there is no evidence that unpublished (but in the public domain) trials in the field of literacy learning are either smaller or have different effect sizes from published studies. However, this still does not tell us whether unpublished studies (not in the public domain) are materially different from those that are publicly available.

- No single review contained large enough numbers of trials to allow confidence in existing techniques for identifying publication bias.
- An important finding of this methodological review is that, because of problems with the presentation of data in most of the reviews, it was possible to draw funnel plots for only three out of a total of 14 systematic reviews.
- Normal quantile plots contributed important information to interpretation of two of the reviews (Ehri *et al*, 2001b; Bus and van IJzendoorn, 1999), which was extra to that obtained from the funnel plots. In both of these studies the normal quantile plot was suggestive of heterogeneity of included studies. This finding is important. If studies with educational heterogeneity are inappropriately placed into a meta-analysis this can produce misleading results. We may erroneously conclude, for example, that an educational intervention is beneficial for a wider population of learners when, in fact, it is only really effective among a subgroup of learners. The use of this simple graphical technique can signal caution to the policy maker that it may be wise to invest in more research before a policy is widely implemented.
- The reappraisal of the US National Reading Panel review (Ehri *et al*, 2001b) found that systematic phonics instruction is associated with an increased improvement in reading accuracy. The effect size was 0.27, which translates into an approximate 12% absolute improvement in a reading accuracy test that is standardised to have a score with a mean of 50% for children not receiving systematic phonics instruction (see Torgerson, 2003, Item 1, p.86). However, this finding still needs to be treated with caution. There was significant heterogeneity within the meta-analysis, which could not be explained by the *observable* (or reported) design characteristics of the included trials or by the educational characteristics of the children included in the studies. Therefore, it

is unclear whether systematic phonics teaching is beneficial to *all* children with different learner characteristics. Only one of the included trials was undertaken in a UK context, which raises concerns about the applicability of the results to a UK setting. In addition, there is the issue of publication bias. The strong possibility of publication bias affecting the results cannot be excluded. Whilst one strength of systematic reviews is their transparency, which should permit easy replication, it was found that this was not a straightforward process. Some of the trials included within the original phonics review did not present data such as the numbers of children randomized to each condition. Therefore, in order to obtain these data authors had to be contacted. One source of heterogeneity was explained by inappropriate pooling of studies. Despite an exhaustive search for unpublished studies, however, there still appeared to be publication bias, and there was significant statistical heterogeneity, which could not easily be explained. Overall, the updated review, whilst broadly confirming the effectiveness of systematic phonics instruction, still requires caution about whether it should be widely implemented. The revised estimated effect size was substantially smaller than that estimated by Ehri *et al* (2001b). Because there was still an issue with publication bias even this estimate of effectiveness may be an overestimate.

Item 3: Design bias

- In terms of the second research question, the individual RCTs that were identified in this portfolio tended to be small and methodologically weak.
- With regard to the third research question, although there were some associations between methodological quality and effect size, and these

differences were at times large, no statistically significant associations were found.

- Although it is considered 'good practice' to undertake blinded follow-up and intention to teach analysis, among other factors affecting internal validity, there was no statistically significant relationship between these methodological factors and effect sizes. This may be because the final sample size of included trials was relatively small, and therefore there was a lack of statistical power. Therefore, there was a real danger of a Type II error.

Item 3: Quality appraisal of systematic reviews

- Replication of a high quality review (Ehri *et al*, 2001b) was not straightforward and led to some discordant results. The search strategy did not seem to exactly replicate the original results. Including some identified trials was difficult as data were missing from the publication. This suggests, therefore, that the measure of assessing quality, i.e., the modified QUORUM guidelines (Shea *et al*, 2001), may not be sufficiently detailed to identify quality issues in reviews that may require a more cautious interpretation of their substantive findings. The more in-depth detailed methodological work in Item 3 modifies the conclusions generated in the earlier part of the portfolio, i.e., Items 1 and 2. A revised conclusion is that although QUORUM is an important first step in evaluating the quality of systematic reviews it is not a 'gold standard' measure of systematic review quality.

In the following section, recommendations flowing from the findings of this portfolio are made.

Recommendations

Design and conduct of randomized trials in literacy learning

- As a general observation, the reporting of RCTs in literacy learning is variable, but generally weak. In the first instance researchers should report sufficient data to allow confirmation and recalculation of the main effect sizes and corresponding confidence intervals. Some trials do not report such basic information as the sample size in each group, how many participants were lost between allocation and follow-up, and the standard deviations of the mean scores at post-test. Journal editors and referees should insist that these minimal data should be included within any report of a RCT.
- In the medium term, educational journals should adopt common guidelines for the reporting of trials, similar to the CONSORT statement used by many healthcare journals.
- In the longer term, the quality of RCTs in literacy learning in particular, and in educational research more generally, needs to be improved. Many trials are too small and may be subject to a Type II error. Many trialists do not explain how sample sizes were derived or how random allocation was undertaken. Furthermore, many trialists either do not use, or do not report the use of, blinded follow-up at post-test. Reporting of confidence intervals is virtually absent.
- In short, the quality of the design, conduct and reporting of randomized controlled trials in literacy research needs to be radically improved. The need for high quality randomized trials is crucial for systematic reviews in this area. Even the best-designed and -conducted systematic review cannot remedy the deficiencies of poorly reported RCTs. Therefore, systematic reviews containing

flawed RCTs may themselves produce biased findings and mislead policy and practice.

- It would seem sensible that work using a larger sample of trials that would give adequate power to see if statistically significant relationships between methodological factors and effect sizes do exist should be carried out.

Conduct of systematic reviews of randomized trials

Most of the systematic reviews identified in this thesis appeared to be of high quality, as judged by the modified QUORUM criteria (although subsequent analysis modified this judgement). One aspect of review quality was the focus of attention in this portfolio, namely publication bias. Few systematic reviews used graphical approaches to assess the potential existence of publication bias. Two graphical methods, funnel and normal quantile plots, were applied to a sample of systematic reviews in literacy learning. None of the systematic reviews had sufficient numbers of trials for funnel plots to be robust. At best they could only be suggestive of publication bias. The normal quantile plots did not appear to add any benefit for the detection of publication bias; nevertheless, their use did point to another potential problem within systematic reviews: potential heterogeneity of the included population of trials.

- It is therefore important that, in future, systematic reviewers in literacy research present data in their reviews to enable others to appraise the individual studies contained in their review.

Future research

- The whole field of literacy learning requires large, rigorous trials to be conducted as a matter of urgency.

- With regard to whether or not the exclusive widespread use of systematic phonics in early literacy development, as recommended in the UK by many stakeholders (policy makers, politicians, teachers), should be implemented, my conclusion is that a more cautious approach is justified. The adoption of this approach could only be fully justified by positive evidence from a large pragmatic RCT within a UK setting. Such a trial would probably be of a cluster design, with schools being allocated to either maintain current adherence to the Primary National Strategy (the control group), or to receive additional systematic phonics instruction or replacement of the PNS by exclusive systematic phonics teaching (the intervention groups).
- An area of further research might be to replicate a sample of systematic reviews. In the present portfolio only one review was replicated, due to resource and time constraints. Therefore, it was not possible to ascertain whether the discordant results from this particular review would be applicable more widely. Therefore, replication of a sample of systematic reviews would be warranted. In the meantime one key recommendation is that, for any review that could potentially induce a change in educational policy, replication by an independent research team is warranted before any such change is initiated.

Key messages

Systematic reviews need to be perceived as an objective guide to decision making. The possibility of meta-analysis being misleading should always be considered, given the possibility of the introduction of bias in the process of including the studies.

- Quality appraisal of systematic reviews using a checklist like the QUORUM guidelines may not reveal significant weaknesses in a review. Therefore, the data extraction and quality appraisal of systematic reviews should be extremely detailed before their conclusions can be relied upon for policy and practice decisions. Replication of some stages of the review is recommended, including as a minimum the replication of data extraction of a sample of included studies and the recalculation of effect sizes. For systematic reviews that have major educational implications, a replication of the review by a second independent research team is warranted, and should be considered as the 'gold standard' for quality appraisal of the review.
- As a minimum, high quality systematic reviews should routinely report the following: detailed search strategy; search for grey literature; effect sizes and means, standard deviations of means and numbers in each condition for each included trial to enable recalculation of effect sizes.
- Examination of systematic reviews for publication bias should be undertaken routinely. An analysis of funnel plots is a useful test for the possible presence of publication bias in a meta-analysis. However, the capacity to detect publication bias is limited when meta-analyses contain small numbers of trials; and interpretation of funnel plots is difficult. Both of these factors should be acknowledged as limitations. The results from normal quantile plots should be used in conjunction with funnel plots to detect possible publication bias.
- For topics of important policy interest, two independent reviewers or teams of researchers should undertake a meta-analysis.

References

Altman, D.G. (1996) Better reporting of randomised controlled trials: The CONSORT statement, *British Medical Journal*, 313: 570-571.

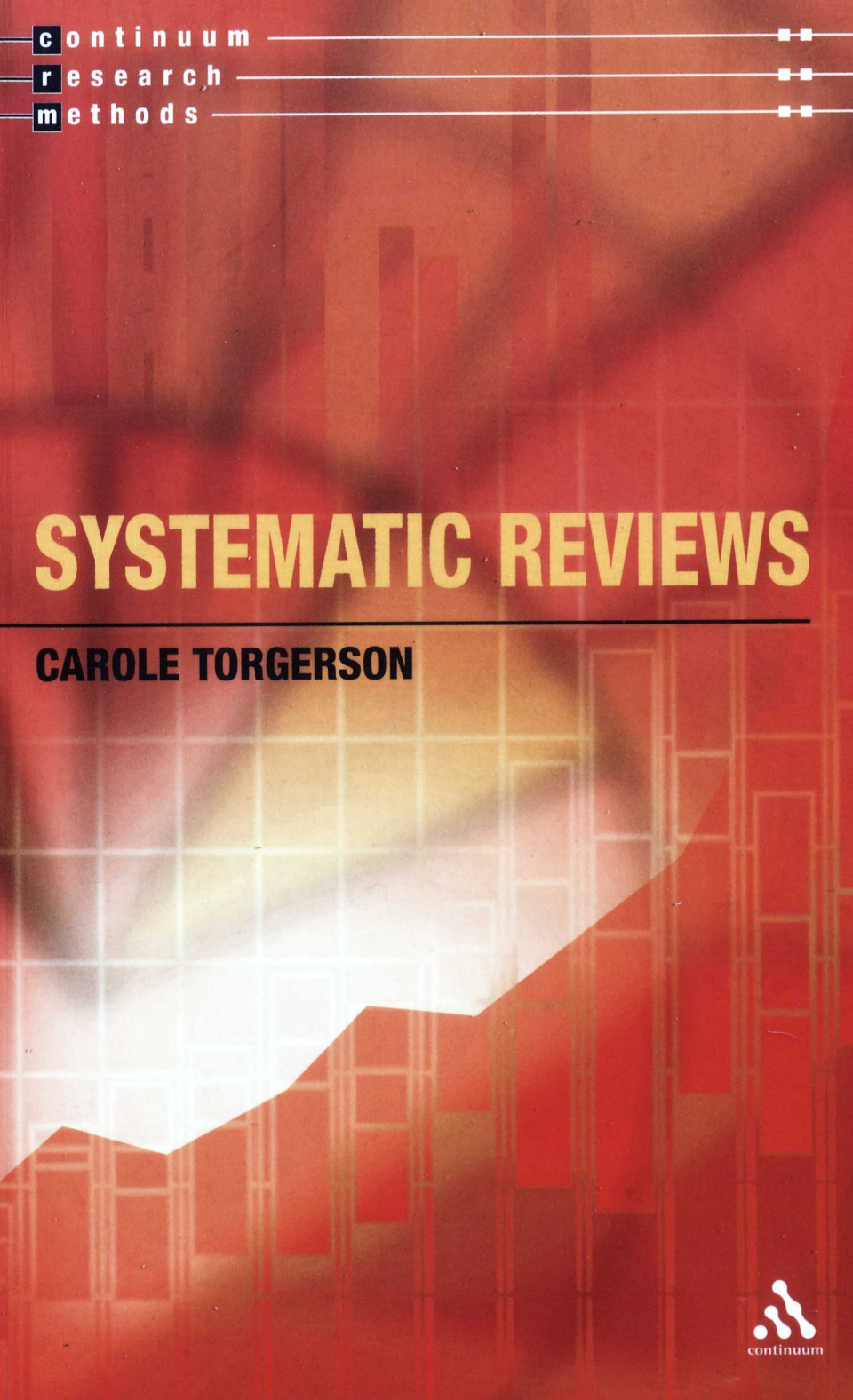
Moher, D., Cook, D.J. Eastwood, S. Olkin, I., Rennie, D. and Stroup, D.F. (1999) Improving the quality of reports of meta-analyses of randomized controlled trials: The QUORUM statement. Quality of reporting of meta-analyses, *Lancet*, 354: 1896-1900.

Shea, B., Dube, C. and Moher, D. (2001), Assessing the quality of reports of systematic reviews: The QUORUM statement compared to other tools, in M. Egger, G. Davey-Smith and D. Altman (eds), *Systematic Reviews in Healthcare: Meta-analysis in Context* (second edition), London: BMJ Publishing Group.

continuum research methods

SYSTEMATIC REVIEWS

CAROLE TORGERSON



continuum

Systematic Reviews

Carole Torgerson

continuum
LONDON • NEW YORK

Continuum International Publishing Group
The Tower Building
11 York Road
London SE1 7NX

15 East 26th Street
New York
NY 10010

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

ISBN 0 8264 6580 3 (paperback)

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress.

Typeset by BookEns Ltd, Royston, Herts.
Printed and bound in Great Britain by MPG Books Ltd, Bodmin, Cornwall

Contents

Series Editor's Introduction

The Continuum Research Methods series aims to provide undergraduate, Masters and research students with accessible and authoritative guides to particular aspects of research methodology. Each title looks specifically at one topic and gives it in-depth treatment, very much in the tradition of the Rediguide series of the 1960s and 1970s.

Such an approach allows students to choose the books that are most appropriate to their own projects, whether they are working on a short dissertation, a medium-length work (15–40 000 words) or a fully-fledged thesis at MPhil or PhD level. Each title includes examples of students' work, clear explication of the principles and practices involved, and summaries of how best to check that your research is on course.

In due course, individual titles will be combined into larger books and, subsequently, into encyclopaedic works for reference.

The series will also be of use to researchers designing funded projects, and to supervisors who wish to recommend in-depth help to their research students.

Richard Andrews

Acknowledgements

Tables 4.3, 4.4 and 5.3, and figure 6.2 were previously published in Torgerson *et al.* (2003), 'A systematic review and meta-analysis of randomised controlled trials evaluating interventions in adult literacy and numeracy', *Journal of Research in Reading*, 26(3), (tables 1, 2 and 5, and figure 1) (copyright-holder UK Reading Association).

Tables 7.1 and 7.2 and figure 7.1 were previously published in Torgerson *et al.* (2002), 'Do volunteers in schools help children learn to read? A systematic review of randomised controlled trials', *Educational Studies*, 28(4), 433–44 (tables I and II and figure 1) (see http://www.tandf.co.uk).

Tables 4.3, 4.4, and 5.3, box 4.3 and figures 6.1 and 6.2 were previously produced in a report to the National Research and Development Centre for Adult Literacy and Numeracy (Torgerson, Brooks *et al.* 2003, *Adult literacy and numeracy interventions and outcomes: expert, scoping and systematic reviews of the trial literature*) (tables 4.1 and 4.2. from 'scoping' review; tables 3.2 and figures 3.1 and 3.2 from 'systematic review of RCTs'; first table in Appendix E).

I acknowledge the influence of the methods of NHS CRO and the EPPI Centre on some of the ideas in the book.

I thank Richard Andrews (University of York) and Greg Brooks (University of Sheffield) for their helpful comments on an earlier version of the manuscript.

I thank David Torgerson (University of York) for suggesting examples and comparisons from health care policy and research, and Alison Robinson (University of York) for proof reading the manuscript.

Scope of the Book

The focus of this book is on systematic reviews and meta-analyses of randomized controlled trials in educational research. Systematic reviews can be used to synthesize studies of other designs, such as longitudinal studies, non-randomized controlled trials or qualitative research (Gough and Elbourne 2002). However, the randomized controlled trial (RCT) is widely acknowledged to be the 'gold standard' of effectiveness research, and systematic reviews of RCTs the gold standard in effectiveness research synthesis. In effectiveness research in education, the randomized controlled trial is the most robust method of establishing 'what works'. Systematic reviews can facilitate the definition of a future research agenda, and inform the evidence-base for policy-making and effective professional practice. This book is aimed at students and researchers who wish to undertake a systematic review for the first time. It includes a step-by-step description of the rationale for and the processes involved in undertaking systematic reviews and meta-analyses.

Glossary

Attrition – Often participants are lost during a trial and cannot be included in the analysis. This is termed attrition or is sometimes known as mortality.

Bias – A term denoting that a known or unknown variable (rather than the intervention) is, or may be, responsible for an observed effect.

Concealed allocation – This is where the researchers, participants and teachers are prevented from knowing in advance the allocation of an individual. Random allocation can be undermined by selection of participants to be in a desired group. Fair randomization will, on average, produce equivalent groups. Using 'open' randomization methods such as random number tables means that the researcher will know the next allocation in advance of it happening. Therefore, in theory the next participant could be 'excluded' from the study if he/she does not possess certain 'desirable' characteristics. This can then lead to bias, which undermines the whole basis of random allocation. It is important, therefore, that the 'mechanics' of randomization are clearly described to assess whether or not the study is susceptible to 'subversion' bias.

Confidence intervals The point estimate of effect of any intervention will always be imprecise. The level of the imprecision is dependent upon the sample size and event rate in the treatment groups. The use of confidence intervals (usually 95%, but sometimes 99% or 90%) reflects this imprecision in the study results. Thus, for example, a treatment that has an effect size of 0.50, but a confidence interval of −0.1 to 1.2 is not statistically significant but will indicate to the reader that there is a relatively high possibility that there is a beneficial effect of treatment in excess of 1 standard deviation. In this instance, one might

consider doing a further, larger randomized trial. In contrast, if the point effect is 0.05 and the confidence interval is −0.1 to 0.12 then the reader might consider that it is unlikely even with a bigger trial that this intervention would show an educationally significant effect (assuming the conduct of the trial in question is of high quality).

CONSORT – Consolidated Standards for Reporting Trials is the methodological standard adopted by many medical journals for publication of randomized controlled trials.

Controlled trial (CT) – This usually means a study with a control group that has been formed by means other than randomization. Consequently the validity of the study using this design is potentially threatened by selection bias.

Co-variates or confounders – These are variables that are associated with outcome. Randomization is the only method that ensures that both known and unknown co-variates are equally distributed among treatment groups.

Effect size – When an outcome variable is measured on a continuous scale (e.g. changes in a test score) the improvement or decrement is described in standard deviation units, which is termed the effect size.

Funnel plot – A method of assessing whether there is any publication bias. The effect size of each study is plotted against its sample size. Small studies will have large random variations in their effect sizes, which will be scattered along the x-axis close to the bottom of the y-axis. Larger studies will be higher up on the y-axis and less scattered along the x-axis. A review with no publication bias will show a plot in the shape of an inverted funnel.

ITT analysis – Intention to Teach – This is where all participants are analysed in their original randomized groups; it is the most robust analytical method. Once participants have been allocated to their respective groups it is important that they remain in those groups for analysis, to avoid bias. A common, but incorrect, method is to exclude some participants after randomization for a variety of reasons. One approach is to do what is termed 'an on-treatment

analysis' – this is where only those participants who demonstrate treatment fidelity are included in the analysis. Unfortunately, this can lead to bias, as those participants who complete treatment are likely to be different from those who do not. Intervention-received analysis can therefore produce a biased result.

Paired randomization – This is a commonly used method in educational research. Participants are formed into matched pairs on the basis of important co-variates (e.g. gender and/or pre-test scores). Once the study group has been formed into pairs a random member of each pair is allocated to the intervention. The consequence of pairing is that there should be exactly equal numbers in each group and the group should be exactly balanced in terms of the characteristics on which the pairing took place. If the co-variate used for pairing (e.g. age) has an unusual relationship with outcome this cannot be explored in the analysis as the pairing eliminates all variation due to that co-variate.

Publication bias – Not all RCTs are published. There is a well-established tendency for trials that produce negative effects or null effects to be less likely to be published than positive trials. Unless a systematic review includes these negative trials it can give a misleading optimistic assessment of the intervention. Existence of publication bias can be detected by using funnel plots.

Randomized Controlled Trial (RCT) – This is where two or more groups have been formed through random allocation (or a similar method). This is the only method that ensures that selection bias is eliminated at baseline.

Regression to the mean – This statistical phenomenon occurs when test results are, by chance, some distance away from the mean. Consequently at post-testing the 'extreme' results will tend to regress to the mean. When selecting participants on extreme test results (e.g. very poor pre-tests) there will be an apparent dramatic improvement on post-test because of this effect (irrespective of the teaching method). Randomization automatically controls for regression to the mean effects. Nevertheless, it can still have an influence if the groups are unbalanced at baseline on pre-test scores. This imbalance can be adjusted for by a multivariate analysis.

Sample size calculations – Trials in educational research commonly exhibit a Type II error. This is where the sample size is insufficient to show, as statistically significant, a difference that is educationally important. Reviews of educational interventions have shown that most interventions will, at best, only lead to an improvement in the region of half a standard deviation and quite often somewhat less. Statistical theory shows that to reliably detect (with 80% power) half a standard deviation difference as statistically significant ($p = 0.05$) for a normally distributed variable requires a minimum sample size of 126 participants. Studies that are smaller than this risk erroneously concluding that there was not a significant difference when actually there was. Therefore, a good-quality study ought to describe the reasoning behind the choice of sample size.

Selection bias – This occurs when groups are formed by a process other than randomization and means that important factors that are associated with outcome differ between the groups *before* they are exposed to the intervention.

Standard deviation – A measure of spread or dispersion of continuous data. A high standard deviation implies that the values are widely scattered relative to the mean value, while a small value implies the converse.

1

Background: Evidence-based Education

Evidence-based policy-making

Since the late 1990s an increasingly high profile of 'evidence-based' policy-making in education and the social sciences has emerged in the UK (Davies 1999, Davies 2000, Constable and Coe 2000, Davies, Laycock *et al.* 2000, Evans and Benefield 2001, Young *et al.* 2002, Gough and Elbourne 2002). The movement towards evidence-based education clearly derives, in part, from similar, earlier developments in health care research. In the early 1990s health care research became dominated by the need to inform policy-making through the use of rigorous evidence from research synthesis. For example, the Cochrane Collaboration was established in Oxford in 1993 (http://www.cochrane.org). Its remit was to undertake systematic reviews of health care interventions through the work of 50 'review groups' in various fields of health care. The National Health Service Centre for Reviews and Dissemination at the University of York (NHS CRD) (http://www.york.ac.uk/inst/crd/srinfo/htm) was established at around the same time to undertake systematic reviews in health care policy. A key element in the recent development of evidence-based education is a renewed focus on systematic review methodology and methods in this field.

A number of reasons have been suggested for the rise of

evidence-based policy, including developments in information technology in general and in electronic databases in particular (Davies, Nutley, Smith 2000). The impetus towards evidence-based education occurred at around the same time as the debate about the value and methods of educational research in the late 1990s (see for example, Hargreaves 1996, Hargreaves 1997, Hammersley 1997, Tooley and Darby 1998). Various criticisms were levelled at the educational research community, most notably for its lack of scientific rigour, quality and relevance. A detailed analysis of the debate is beyond the scope of this book. It has been well documented elsewhere (Sebba in Davies, Laycock *et al.* 2000, Davies 2000, Pring 2000, Evans and Benefield 2001, Pirrie 2001, Oakley 2002). However, links between the debate and the rise of the evidence-based education movement have been suggested (Pirrie 2001, Oakley 2002). The trend towards 'evidence-based' and then 'evidence-informed' policy extended to spheres in the social sciences, including health promotion in the early 1990s, and education in the late 1990s and into the beginning of the twenty-first century. Recent significant national and international developments in evidence-based educational and social research have raised the profile of the movement.

In 1997 a series of biennial conferences entitled 'Evidence-based Policies and Indicator Systems' (http://cem.dur.ac.uk/ebeuk) was established at the University of Durham. At the time of the first conference in 1997 the concept of 'evidence-based' policy outside the field of health care was still fairly 'marginal' (Constable and Coe 2000).

Recently, two evidence-based initiatives have been established in the UK to play prominent roles in the 'evidence-based policy' (EBP) debate. The Centre for Evidence-Based Policy and Practice was established in December 2000 (funded by the Economic and Social Research Council) at Queen Mary, University of London (http://www.evidencenetwork.org), to 'advance' the debate

about evidence-based policy and practice (see Young *et al.* 2002).

The Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre) (http://www.ioe.ac.uk/projects.html) within the Social Science Research Unit at the Institute of Education, University of London, began undertaking systematic reviews in health promotion (funded by the Department of Health) in 1993. In 2000 the Department for Education and Skills (DfES) funded the Centre to support a series of systematic reviews in educational research (http://eppi.ioe.ac.uk/EPPIWeb/home.aspx), to inform policy, practice and 'democratic debate' (Gough and Elbourne 2002). This initiative is based on the Cochrane Collaboration model: a number of collaborative review groups have been set up to undertake systematic reviews in various areas of educational research. (For detailed analyses of the work of the EPPI-Centre see Sebba in Davies, Laycock *et al.* 2000, Evans and Benefield 2001, Gough and Elbourne 2002, Oakley 2002.) Another UK initiative occurred in 2003: the Teacher Training Agency commissioned a series of systematic reviews, relevant to initial teacher training and supported by the EPPI-Centre.

Another important recent development in social and educational research has been the establishing of the Campbell Collaboration Social, Psychological, Educational and Criminological Trials Register (C2-SPECTR) in February 2000 in the USA. Its aims are to identify all the experimental research of educational, social policy and criminal justice interventions (http://campbell.gse.upenn.edu/), and to undertake, update and make accessible systematic reviews of social and educational interventions (Petrosino *et al.* 2000). The Campbell Collaboration mirrors the earlier Cochrane Collaboration. Indeed, the C2-SPECTR initiative arose directly out of the work of the Cochrane Developmental, Psychosocial and Learning

Problems Group, and was influenced by the methodology and methods developed in the Cochrane Collaboration (Petrosino *et al.* 2000).

The need for reviews

These initiatives and others developed from the realization that a single experiment seeking to investigate the effectiveness of an educational policy, no matter how well conducted, is limited by 'time-, sample- and context-specificity' (Davies 2000). A single study should not be considered in isolation, but positioned within the 'totality' of research in a field to give a more complete picture (Mulrow 1994, Chalmers *et al.* 2002).

Government policy in the UK has not always taken cognizance of systematic reviews, sometimes preferring the results of single studies. In 2000 the government introduced a 'driver education' programme on the basis of positive results from a single study (Clayton *et al.* 1998). Yet a systematic review of trials found that the introduction of driver education into schools actually led to an *increase* in injuries and road traffic accidents among younger drivers (Cochrane Injuries Group Driver Education Review 2001). This counter-intuitive finding was explained by the fact that young people who had participated in the driver education programmes tended to start to learn to drive earlier compared with young people who had attended control schools and therefore not participated in the programme.

There are other problems with single studies. The results of an educational experiment undertaken in the USA might not be generalizable to the UK educational context. However, if a series of high-quality experiments have been undertaken recently in the USA, Canada, Australia and New Zealand and if they are sufficiently similar in terms of sample and context, and are found to yield similar results in

a systematic research synthesis, one might, in the absence of any UK evidence, be confident that the effects could transfer to a UK setting.

Traditional narrative literature review

The traditional literature review often forms the basis of 'opinion' pieces, 'expert' reviews or students' theses, but is less helpful for guiding policy or contributing to an informed debate of the issues. This is due to a number of factors. The research literature included in traditional narrative reviews tends to be a 'biased' sample of the full range of the literature on the subject. It is usually undertaken through the perspective of the reviewer who gathers and interprets the literature in a given field. The reasons for including some studies and excluding others are often not made explicit, and may reflect the biases of the author. Included references may be used to support the 'expert opinion' whilst other references that contradict this opinion may be excluded from the review. If the search strategy and inclusion criteria have not been made explicit it will not be possible for the review to be replicated by a third party. Because a 'systematic, rigorous and exhaustive' search of all the relevant literature (Davies 2000) has not been undertaken, relevant studies could have been excluded from the review, leading to potential selection and/or publication bias. Failure to include all the appropriate studies may lead to incorrect interpretations of the evidence. Finally, usually in traditional literature reviews the individual studies are not quality assessed before inclusion in the review, and therefore there may be no differentiation between methodologically 'sound' and 'unsound' studies (Young *et al.* 2002). The shortcomings of the 'expert' review are well recognized in evidence-based health care, where in the hierarchy of evidence 'expert opinion' constitutes the lowest

grade of evidence. Nevertheless, they remain influential in all areas of social research.

Systematic review

A more rigorous alternative to the 'narrative' review is the systematic review. A systematic review differs from a traditional narrative review in that its methods are explicit and open to scrutiny. It seeks to identify *all* the available evidence with respect to a given theme. Systematic reviews have the advantage of including all the studies in a field (sometimes positive and negative studies), so the reader can judge using the *totality* of evidence whether the evidence supports or refutes a given hypothesis. This evidence is collected, screened for quality and synthesized into an overall summary of the research in the field. Because all the evidence pertaining to a given topic is included in the systematic review, with rejected evidence catalogued and the reasons for rejection made explicit, the resulting findings are often less susceptible to selection, publication and other biases than those of a traditional or 'non-systematic' review.

Rationale for systematic reviews

The rationale for undertaking a systematic review has been well rehearsed in the fields of health care (for example, Egger *et al.* 2001, Chalmers *et al.* 2002), social and educational research (for example, Evans and Benefield 2001, Oakley 2002). It is a scientifically rigorous method for summarizing the results of primary research and for checking consistency among such studies (Petticrew 2001). The rationale for systematic reviews in medicine is firmly embedded in the 'positivist' or 'scientific' tradition or paradigm (Mulrow 1994). Systematic reviews are traditionally associated with

meta-analyses of research based on quantitative epistemological traditions and methodologies (Badger *et al.* 2000, Hammersley 2001). The most important aspect of the scientific paradigm is that a study must be replicable as well as reliable and credible. Judgements and assumptions are made explicit to allow exposure to scrutiny and comment. In education (and health) the methodology is currently being extended to encompass studies using a broad range of methods including qualitative research (Petticrew 2001).

Mulrow (1994), writing about systematic reviews and meta-analyses in the field of health care, has outlined some assumptions on which this rationale is based. Systematic review methodology has the ability to manage potentially 'unmanageable amounts of information', and rationalize existing evidence efficiently by establishing whether research findings are consistent and generalizable, and to explain why if they are not. Often similar studies can be put together, statistically, in a meta-analysis. Meta-analyses can be used to increase power and precision in the measurement of effect sizes. Finally, systematic reviews use scientific methods that reduce 'random and systematic errors of bias' (Mulrow 1994). Clearly such a rationale does not apply exclusively to research in health care.

Aims of a systematic review

The aims of a systematic review are well documented in the health care (Mulrow, 1994, Egger *et al.* 2001, Petticrew 2001, Chalmers *et al.* 2002) and social policy literature (Davies 1999, Davies 2000, Badger *et al.* 2000, Gough and Elbourne 2002, Young *et al.* 2002). They are:

(i) to address a specific (well focused, relevant) question;
(ii) to search for, locate and collate the results of the research in a systematic way;

(iii) to reduce bias at all stages of the review (publication, selection and other forms of bias);
(iv) to appraise the quality of the research in the light of the research question;
(v) to synthesize the results of the review in an explicit way;
(vi) to make the knowledge base more accessible;
(vii) to identify gaps; to place new proposals in the context of existing knowledge;
(viii) to propose a future research agenda; to make recommendations;
(ix) to present all stages of the review in the final report to enable critical appraisal and replication.

Definition and origin of systematic reviews

Research synthesis is *secondary research*. It involves techniques and strategies to accumulate the findings of *primary research* (Davies, Nutley, Smith 2000). Systematic reviews and meta-analyses are key features of research synthesis. They involve the search for, location, quality appraisal and narrative synthesis of all the relevant studies in a field. Meta-analysis involves the computation of a combined 'effect size' across the studies in a given field (Davies, Nutley, Smith 2000). Useful definitions of 'systematic review' and 'meta-analysis' are quoted in Chalmers *et al.* 2002:

> SYSTEMATIC REVIEW The application of strategies that limit bias in the assembly, critical appraisal, and synthesis of all relevant studies on a specific topic. Meta-analysis may be, but is not necessarily, used as part of this process. (pp. 176-7)
>
> META-ANALYSIS The statistical synthesis of the data from separate but similar, i.e. comparable studies, leading to a quantitative summary of the pooled results. (p. 14) (Last, 2001, *Dictionary of Epidemiology*, quoted in Chalmers *et al.* 2002)

There is a long history of the use of systematic review techniques in educational research to search for, retrieve and synthesize literature in a number of different contexts (Slavin 1986, Lipsey and Wilson 1993, Davies 2000, Petrosino *et al.* 2000). Educational researchers were among the early users of systematic reviews, although astronomers claim to be the first group of researchers to use the method (Slavin 1986, Petticrew 2001). Recently, Chalmers *et al.* (2002) have described a 'long history' of research synthesis in various disciplines and diverse subjects: medicine (i.e. treatments for scurvy in the eighteenth century), agriculture (also the eighteenth century), astronomy, and the 'psychology of time' (nineteenth century). The 'science of research synthesis' emerged in the twentieth century with Pearson's 1904 review of evidence on the effects of a vaccine against typhoid and some early reviews in the field of educational research (Chalmers *et al.* 2002). Educational researchers started combining the results of educational experiments in the first half of the twentieth century. In the 1950s and 1960s social science researchers explored different statistical approaches for undertaking meta-analyses. This was particularly so in the fields of education and psychology (Chalmers *et al.* 2002). Social scientists published many texts in the 1970s and 1980s on statistical approaches to meta-analysis and data synthesis (Glass 1976, Lipsey and Wilson 2001). It was only in relatively recent times that 'modern' health care researchers realized the merits of undertaking systematic reviews – although, as Chalmers *et al.* (2000) point out, James Lind performed a systematic review on treatments for scurvy in the eighteenth century. Certainly, it is only since the mid-1980s that there has been an explosion of systematic reviewing in health care.

Although educational researchers have undertaken 'research synthesis' for some considerable time (Slavin 1986) the term 'systematic review', however, is a relatively recent one and was initially used in health care research.

Traditionally, educational reviewers have used terms such as 'meta-analysis' and/or 'research synthesis' (Slavin 1986). The term 'meta-analysis' has a specific statistical meaning. It is a method of combining a number of quantitative studies in order to produce a more precise estimate of effect than can be achieved by any single study. It is not always the case that meta-analyses are undertaken on studies that have been identified using systematic review methods. Often it is not possible to combine studies statistically, and the synthesis of the findings of a range of quantitative studies can be done in other ways. Even when studies can be statistically combined this may not be appropriate. Slavin (1986) criticized educational researchers for combining all quantitative studies into a meta-analysis without due regard to their quality or context. He argued for a 'best-evidence synthesis' approach, whereby studies of certain quality and or contextual criteria are included within a meta-analysis. While the term 'research synthesis' moves the reviewer away from the requirement to undertake a meta-analysis of all identified studies it still implies some form of 'pooling' or meta-analysis of the data extracted from identified studies. Often this is simply not possible. Some studies, for example, may report no outcome data, which precludes their inclusion within a meta-analysis. Nevertheless, identification and reporting of the existence of these studies can be valuable, even if their results cannot be synthesized in quantitative ways. The systematic review process differs from meta-analysis and research synthesis in that it describes the whole process of identifying all the relevant literature within a given area.

Criticisms of systematic reviews

There have been a number of attacks on the rationale of systematic reviews (see for example, Pirrie 2001). The

methodology of systematic reviews has been criticized because it is founded on 'questionable' premises about the nature of reviewing and ideas about research (see in particular, Hammersley 2001). Eysenck (1995) has criticized the 'mechanical' nature of the review process without sufficient regard to the quality and interpretation of the data.

There has been scepticism about the utility of systematic review methodology outside reviews of health care interventions using quantitative research designs. Conversely the relevance of the 'medical model' of research synthesis (based on the randomized controlled trial) to education has been questioned (Constable and Coe 2000, Pring 2000, Evans and Benefield 2001, Hammersley 2001).

Whilst many of the criticisms of systematic reviews have some merit, the alternative to a systematic review is bleaker: a narrative review with selection of studies based on the possible biases of the author. A more helpful approach to criticizing the conduct of systematic reviews is provided by Slavin (1986). Slavin has made similar points to Eysenck, in suggesting that some meta-analyses include all manner of quantitative studies – the 'good, bad and indifferent' (Eysenck 1995). Nevertheless, Slavin argues that the best-evidence synthesis approach (systematic review) that includes some aspect of quality appraisal moves us away from the unthinking mechanical nature of meta-analysis criticized by Eysenck, although this approach does reintroduce an element of 'subjectivity'.

Systematic reviewing is not 'value free'

Although the methods of a systematic review are firmly based within the 'positivist' tradition it is not a 'value free' approach. Early educational meta-analyses sought to avoid imposing the values of the researcher on the evidence by including all quantitative data in the review on the basis

that bias can be introduced through the process of choosing to exclude certain studies (Slavin 1986). This procedure, however, can introduce bias by allowing poor-quality studies to influence the outcome. Slavin (1986) cogently argued that some form of judgement must be used to avoid introducing bias into the review from irrelevant or poor-quality studies. Thus, features of the systematic review method include value judgements throughout the process. These values include an explicit statement of the nature of the data that will be extracted from the included papers and details of the assumptions underpinning the basis on which the reviewer will interpret the included papers. The reviewer's values are implicit at each stage of the review, from the initial searches until the final synthesis. They are implicit in the choice of key terms employed in the searches, in the inclusion and exclusion criteria selected by the reviewer, and they are present in the interpretation of the included papers. Because the procedures in a systematic review are explicit and transparent the values used to inform the review are open to criticism, comment and consequent change by other reviewers. A systematic review of trials in health care suggested that routine administration of human albumin after large amounts of fluid loss, usually due to burn injuries, was associated with *increased* mortality (Cochrane Injuries Group Albumin Reviewers 1998, Roberts 2000). The findings of this review were contested by Wilkes and Navickis, who researched the literature using different inclusion criteria and did not find the alarming increase in mortality observed by Roberts and colleagues (Wilkes and Navickis 2001). Gough and Elbourne (2002) have characterized the process of undertaking a review as 'interpretive' rather than 'mechanical'. Whilst a systematic review approach can reduce some biases it is not, therefore, a value-free enterprise.

Systematic reviews for an 'evidence-informed society'

It is important that policy-makers, teachers and other educational stakeholders have access to the full range of 'evidence' on a subject in order to engage in the 'democratic debate' (Gough and Elbourne 2002) or participate in an 'evidence-informed society' (Young *et al.* 2002). This is the crux of the significance of systematic reviews: they can inform the development of a dynamic relationship between research, policy and practice. Research synthesis can also illuminate the field for future researchers by highlighting the problems of undertaking a study in a particular context. A systematic review may identify any existing relevant high-quality research, which might render redundant the requirement for more primary research (although sometimes research may be required to define the optimum dissemination and implementation strategies). A systematic review can help to inform the design of proposed research studies by giving estimates of effect sizes, which can be used to inform design issues such as sample size calculations. For example, a recent trial randomized 145 students on the basis that this would detect a difference of 0.43 of an effect size (Fukkink 2002); this difference was based on the findings of a previous systematic review. A systematic review can also help to inform the research question.

An interesting example of how systematic reviews informed both research and policy relates to the effect of class sizes on educational outcome. Meta-analyses of early experimental studies on the effect of class sizes on educational achievement indicated a small, but beneficial, effect of reducing class size on measures of achievement (Hedges 2000). These data, however, were criticized from a number of aspects. The studies tended to be small, localized and short-term, and whether the beneficial effect of reducing class size could be sustained to the general school

population remained in doubt. In the 1990s a very large randomized trial undertaken in the state of Tennessee (USA) finally resolved the issue of class size. Children and teachers from 79 elementary schools were randomized to either be taught in or teach classes with a size of about fifteen children compared with classes of 25. The results of the trial confirmed both the beneficial effect of reducing class size and also that this benefit was sustained for several years (Hedges 2000). The results from this very large experiment were very similar to the earlier meta-analyses of small trials. This is an unusual example, in education, of the results of a systematic review of small trials being confirmed by the 'gold-standard' method: the large or mega trial.

Systematic reviews can also examine the external validity or generalizability of randomized trials. Many trials are undertaken in settings that are unlike 'normal' educational practice, for example within a university psychology department with psychologists or educational researchers delivering the intervention. Whether the effectiveness of such an intervention is applicable to a routine educational setting is open to debate. Fukkink (2002) observed, in a re-analysis of a systematic review, that an intervention delivered by researchers tended to have larger effects than when it was delivered in an ordinary school setting. By identifying all the relevant studies a systematic review can compensate for the poor external validity of a single experiment.

Effectiveness research

Both the Cochrane and Campbell Collaborations focus their attention, primarily, on identifying and synthesizing experimental and quasi-experimental research (Petrosino *et al.* 2000) because their priority is on issues of 'effectiveness', i.e. what works in interventions in health care, social policy, criminal justice and education. Experimental

research is the most appropriate evidence of effectiveness in social research (Oakley 2000), and systematic reviews of randomized controlled trials and controlled trials are necessary in order to establish the effectiveness of an educational intervention (Petrosino *et al.* 2000, Gough and Elbourne 2002). In Chapter 2, the importance of experimental methods in educational research and the development of what has become by consensus the 'gold-standard' experimental method, the randomized controlled trial (RCT), are considered.

Summary

- There is an increasingly high profile of 'evidence-based' policy in education and the social sciences.
- A key element of this development is a renewed focus on the rationale and techniques of systematic reviewing.
- The stages of a systematic review aim to limit potential selection, publication and other biases.
- Systematic reviews can inform policy, practice and research.

2

The Nature of Evidence in Effectiveness Research

Educational research in the UK in the last 30 years has been dominated by the qualitative paradigm. Whilst qualitative research can give us important clues as to how and why something may or may not work, can prefigure and clarify issues in a particular field of enquiry and, on occasions, be a useful supplement to quantitative research, it cannot tell us whether or not an educational intervention is actually effective. This requires testing using an experimental or quasi-experimental research design where at least two groups are compared: one receiving the educational intervention under evaluation, the other acting as a control.

Controlled evaluations have been used, intermittently, for many centuries. Chalmers reports several examples of controlled trials of medical treatments, including the famous Lind study of citrus fruits for the treatment of scurvy in the mid-eighteenth century and an evaluation of bleeding among sick soldiers in the Peninsular War at the beginning of the nineteenth century (Chalmers *et al.* 2002). Controlled evaluations increased substantially in the twentieth century. The medical community celebrates the 1948 MRC Streptomycin trial that is generally believed to be the first 'true' randomized trial. Oakley claims that educational and psychological randomized trials pre-date the Streptomycin trial by up to 50 years (1998, 2000). Chalmers, however, whilst acknowledging the existence of relatively robust

quasi-randomized experiments, has found no strong evidence that a 'true' controlled trial using random allocation was performed before the 1948 MRC Streptomycin trial (Chalmers, personal communication, 2001). Nevertheless, there are some important 'classical' quasi-experimental studies undertaken before 1948 that underpin the need for rigorous experimental methodology.

The Cambridge-Somerfield experiment (Oakley 2000) undertaken in 1937 allocated 160 'delinquent' boys to either receive social worker support or to act as controls. Teachers were asked to identify boys at risk, and about 160 boys were randomized to be controls or to receive extended pastoral care. The results indicated no evidence of benefit of the intervention; indeed, there was evidence of harm. In the last follow-up in 1975 42 per cent of the boys in the intervention group had had an undesirable outcome (e.g. criminal conviction, early death, etc.) compared with 32 per cent of the boys in the control group. This was a statistically significant difference. The results of this trial appeared to demonstrate a harmful effect of social workers for such boys. In this instance, an experiment demonstrated that an intervention that was widely believed to 'help' children was actually harmful.

More recently a systematic review of experimental studies of 'scared straight', a method used widely in the USA and, occasionally, in the UK, showed that this intervention to prevent adolescents from turning to crime actually increased criminal activity (Petrosino *et al.* 2002). The dissemination of 'scared straight' was based upon anecdotal evidence that taking juvenile petty offenders to be 'scared' by hardened incarcerated criminals would persuade them not to reoffend. A systematic review of randomized controlled trials, however, showed it had the opposite effect.

In the early 1970s US educational researchers persuaded head teachers and the Federal Government to allow schools with large numbers of African-American pupils, from poor

areas, to be randomized to either receive an added financial injection or to act as a control (Crain and York 1976). There was opposition to the experiment, partly to enable grounds that the amount of money was too small in order a difference to be demonstrated. However, because randomization controlled for confounding factors, the experiment *did* show that the children's test scores significantly improved compared with similar children in control schools. This result persuaded policy-makers to offer similar financial packages to other schools in deprived areas.

In the 1990s a large field trial of small class size was undertaken in Tennessee (costing approximately $12 million) showing the benefit of smaller class sizes (Hedges 2000).

Many of the early educational experiments did not use 'true' randomization. Rather, they used quasi-random methods to form groups. A common method is alternation. Alternation might take the form of allocating all children with a surname starting with A to one group whilst all children with a surname beginning with B are allocated to another group. This method, however, can lead to the formation of biased groups and is generally now being replaced by true random allocation.

There have been numerous RCTs in social and educational research (Petrosino *et al.* 2000), but this research paradigm has received less attention in the past 30 years partly as it has appeared to be the 'loser' in the 30-year paradigm war (Oakley 1998). Recently it has come to the attention of policy-makers that inferences about what does and does not work in education cannot be drawn either from qualitative research or quantitative research of an observational nature. Boruch (1994) notes, in the USA at least, an increase in the use of RCTs in the social sciences, and looks forward to the day when, if at all possible, social science interventions will be evaluated using the RCT. Absence of controlled tests in an area leads to a debate among the 'ignorant' (Boruch 1994).

Randomized controlled trials

The most robust method of assessing whether something is effective or not is the randomized controlled trial (RCT). Cook (2002) has argued that in health and social policy the RCT has long been recognized as the most rigorous method of estimating effectiveness. As far back as the 1920s educational researchers were describing how to experiment in education using a randomly formed control group (Oakley 1998). As with systematic reviews, educational researchers were in the forefront of the design of controlled experiments. More recently, many educational researchers have viewed the RCT as a design that is either not possible to undertake in educational research, or is from an inappropriate paradigm. It is not the intention to explore these issues in this book: they have been well documented elsewhere (see for example, Oakley 2000).

In an RCT participants are randomly allocated to the interventions being evaluated. Typically, a participant will be allocated either to the new intervention (the so-called experimental group) or allocated to whatever is the usual practice (the control group). There are many variants to this design, for example allocating the participants to receive the new intervention either straight away or later (a waiting list design), or to receive both the new and the old intervention but in different randomized sequences (reversal or cross-over design), or allocating groups (in educational research this is usually intact classes or schools) in a cluster design. However, the essence of this design and all its variants is the *random* allocation. If participants are allocated on any other basis, one cannot be sure whether (except for chance differences) the experimental and control groups were similar before receiving (or not receiving) the intervention, and therefore it becomes impossible to disentangle the effects of the intervention from the characteristics of the people being allocated to the

intervention. Techniques can be used to attempt to control for the potential confounding from known variables, but they cannot adjust for unknown variables.

Why randomization?

There are a number of methods of assembling two or more groups for the purposes of comparing whether an intervention is effective or not. The benefits of using random allocation have been described previously (e.g. Cook and Campbell 1979, Torgerson and Torgerson 2001, Torgerson and Torgerson 2003a). It is not proposed to detail the design and strengths of RCTs; however, it is important to discuss their main features in order to be able to distinguish between high- and low-quality RCTs. The two main reasons for using random allocation are to avoid regression to the mean effects and selection bias. Randomization avoids both of these problems; however, selection bias in particular can be introduced after random allocation in poor-quality trials.

Regression to the mean

Regression to the mean (RTM) is a highly prevalent problem that affects most areas of human endeavour. In education it is a particularly severe problem affecting areas such as students' test scores and school league tables. The phenomenon occurs when a variable is measured on one occasion and then is remeasured subsequently. This phenomenon explains why researchers consider the pre- and post-test 'quasi-experiment' the weakest research design (Cook and Campbell 1979). In a pre- and post-test evaluation typically students are selected who have scores that are below some threshold: that is they are scoring

badly. If we retest such students then the 'average' mark will move upwards irrespective of any intervention. If there is an intervention the improvement due to the regression to the mean phenomenon may be erroneously ascribed to the intervention. A review of pre- and post-test studies by Lipsey and Wilson (1993) showed that such studies produce effect sizes with an average 61 per cent greater improvement than studies that use a control group. Part of this exaggerated benefit is almost certainly due to regression to the mean. Forming comparison groups using random allocation deals with regression to the mean as it affects both groups equally and the effect is 'cancelled out' in the comparison between the post-test means.

Selection bias

Selection bias occurs when the groups formed for comparison have not been created through random allocation and are different in some way that can affect outcome. Schools who volunteer to pilot a change in the curriculum will often be different from schools that do not volunteer. Comparing these two groups of schools will be susceptible to 'confounding' as there is likely to be some characteristic present in one group of schools that could explain differences in outcomes. Similarly, individuals who volunteer or ask to take part in an intervention are likely to differ in some way from those who do not. Such differences could again explain any differences in outcome.

Systematic reviews of randomized trials

In an ideal world, evidence-based policy and practice in education for questions of effectiveness should be informed through systematic reviews of the results of large,

well-conducted, randomized controlled trials. In areas where there are RCTs these often tend to be relatively small. Which reduces the possibility of an individual trial giving a clear and unambiguous answer. Systematic review methods are particularly valuable when the field of inquiry contains large numbers of relatively small randomized trials as is so often the case in educational research. When small randomized trials are examined on an individual basis they can give misleading results. This is because they have relatively low statistical power to detect modest but important differences in educationally important outcomes. By using meta-analytical methods similar trials can be pooled to enable the analyst to observe, as statistically significant, worthwhile effect sizes that individual trials may have missed.

Whilst meta-analysis can go some way towards addressing the problem of underpowered trials, it will not produce a true estimate of effectiveness if the trials contained within the analysis are methodologically flawed. In addition, meta-analyses may give unduly optimistic results if there is substantial publication bias, that is if studies that show either a null or negative effect remain unpublished and therefore cannot be included in any form of review. The issue of trials with poor or flawed methodology can be addressed in a systematic review through the use of inclusion and or exclusion criteria. In terms of the problems of publication bias, if the studies cannot be identified then they cannot be included, no matter how exhaustive the review. There are techniques, however, that can be used to identify whether or not for a given review there is a danger of publication bias. This problem and the issue of trial quality will be addressed later.

Because the need to know whether something works or not should be the overarching aim of any body of research, this book makes no apology for the decision to focus solely on procedures for undertaking systematic reviews of randomized controlled trials.

Summary

- Experimental research is essential in questions of effectiveness.
- There is a 'long history' of experimental research in education and the social sciences.
- The most robust method of assessing effectiveness is the randomized controlled trial (RCT).
- Systematic reviews and meta-analyses of RCTs are valuable research tools.

3

The Stages of a Systematic Review

The stages of a systematic review are well established in health care (for example, Egger and Davey-Smith 2001, NHS Centre for Reviews and Dissemination 2001), in social policy (Oakley 2002) and in education (for example, Badger *et al.* 2000, Evans and Benefield 2001):

(i) A *protocol* or plan of the research is written to establish: the theoretical, empirical and conceptual background to the review; the research question(s); the objectives; the scope of the review and the methods for searching, screening, data extraction, quality appraisal and synthesis.

(ii) Within the protocol a set of predetermined written *inclusion and exclusion criteria* are specified. For example, the protocol may specify that only studies employing a 'true experimental' design and written in the English language will be included.

(iii) Once the protocol has been developed and, ideally, peer reviewed, the *literature search* can commence, starting with an electronic search. The literature search may also include hand searching of key journals and other methods of retrieval. The results of the search are then *screened* by at least two independent reviewers, firstly on the basis of titles and abstracts (first stage screening), and secondly on the basis of full papers (second stage screening).

(iv) At the '*scoping*' or '*mapping*' stage the studies retrieved for the review are described and classified. At this stage all of these studies may be data extracted for inclusion in the *in-depth review*, or it may be decided to further refine the research question and inclusion criteria and select a more narrowly focused area for the full systematic review.

(v) Once relevant papers have been identified their data need to be extracted, using a standard data extraction sheet, again using at least two independent researchers (*double data extraction*). Also the studies are assessed to determine their quality (*quality appraisal*). This is usually based on internal validity, but includes some analysis of external validity.

(vi) Extracted data are then summarized in a *synthesis*. This can be done as a 'qualitative' overview if the data are not in a form that permits a statistical summary. If the data are numerical and are of sufficient homogeneity then they can be combined within a *meta-analysis*, which will give an overall figure for the effect of an intervention.

(vii) Finally, the synthesized data will be interpreted within a *report*, which should be exposed to peer-review before publication.

Summary

- The seven main stages of a systematic review are well established in health care, social policy and educational research.
- The stages include: writing the protocol (including the inclusion and exclusion criteria); searching and screening; 'scoping' or 'mapping' the research; extracting data from the included studies and quality appraising them; synthesizing the studies in a narrative, and sometimes in a meta-analysis; writing and disseminating the report.

4

Developing a Protocol; Searching and Screening; Data Extraction

Developing a protocol

The first stage in a systematic review is the development of the review protocol. The protocol is an *a priori* statement of the aims and methods of the review. The idea behind writing a review protocol is that the research question(s), the aims and the methods of the review are considered in advance of identifying the relevant literature. This allows the reviewer to conduct the review with minimal bias, and ensures greater efficiency in the review process. Stating a clear research question before the literature search is undertaken will prevent unnecessary effort and cost in identifying and retrieving irrelevant papers. The criteria for including papers in the systematic review are established *a priori*, in order to reduce the possibility of reviewer selection and inclusion bias, by avoiding the situation where criteria are changed as the review progresses or decisions are made to include studies on the basis of their results. If the decisions are explicit this enables them to be justified. The rationale for developing the protocol as independently as possible from the literature is that this avoids the research question and the inclusion/exclusion criteria being unduly influenced by one or two studies, which can lead to bias. For example, if reviewers are aware of the existence of 'seminal' studies in

the area, they may develop their research question and inclusion criteria to ensure that these particular studies are included. The known studies, for example, may have used non-random allocation processes and would have been excluded if the inclusion criteria specified randomized trials. The reviewers, however, may be tempted to specify inclusion criteria that include 'quasi-random' and controlled trials as well as randomized trials, simply in order to include the known studies.

Previous guidance for systematic reviews in health care can be contradictory. For example, Egger and Davey-Smith (2001), whilst recommending the writing of an *a priori* review protocol, also state that:

> The review protocol should ideally be conceived by a group of reviewers with expertise both in the *content area* and the science of research synthesis. (Egger and Davey-Smith 2001, emphasis added)

Involving content experts in the review will necessarily include viewpoints already 'influenced' by non-systematic knowledge of existing research studies. Therefore, it is often not possible for a researcher to be truly unfamiliar with all the relevant studies. Knowledge of at least some of the existing studies is almost certainly going to influence the protocol. Indeed, it could be argued that a preliminary literature review should be used to influence and refine the protocol. A cursory electronic search (*rapid scope*) can be used to estimate the size of the relevant literature. For instance, a rapid scope of the literature on the effects on reading and spelling of interventions to increase phonemic awareness would reveal large numbers (50 to 60) of experimental studies conducted in the USA in the last 20 to 30 years (Ehri, Nunes, Stahl, Willows 2001). If the review resources are scarce (e.g. a student undertaking a review for a thesis) then it might not be possible to review all the

relevant literature in this vast field. An alternative approach might be to develop a protocol that defines a very narrow research question (e.g. the effect of phonemic awareness training on the development of beginning reading, in 'normally achieving' children aged 4 to 6: this is a relatively small literature).

A scoping review is also important in order to identify existing systematic reviews in the area. If the scoping review uncovers a recent, rigorous review within the area that already addresses the proposed research question it would be unnecessary to repeat the review. In this situation the research question could be refined to address another policy relevant question.

Expert knowledge of a number of existing studies, particularly if they tend to be obscure references, may be helpful in developing the electronic search strategy. An exhaustive strategy ought to identify all known, relevant, papers as well as ones that are unknown. Prior knowledge of an area, therefore, can aid the review process, although the reviewer must be aware that it can also introduce bias to the review. The latter problem can be reduced if clear, consistent and logical justifications are made for the inclusion criteria.

Conceptual issues central to the review should be firmly embedded in the protocol. For example, in systematic reviews of literacy learning the relevant conceptual issues might include the nature of literacy, learner characteristics, literacy outcomes or measures and the nature of the interventions. The conceptual issues will help the reviewers to refine their research question for the review to make it well focused and relevant.

In a high-quality systematic review the inclusion and exclusion criteria are rigorously and transparently reported. The inclusion and exclusion criteria will include the time span of publications, the type of research to be reviewed (study design) and the relevance to the research question.

The protocol will also include criteria for quality appraising the included studies (e.g. CONSORT guidelines, Altman 1996), and the categories for data extraction will be specified. In addition, it can be stated that in the eventual synthesis of the research, more weight will be given to the studies assessed as being of 'higher quality'.

The key features of the application of the inclusion and exclusion criteria are:

- they are established *a priori*;
- they are explicit;
- they are applied stringently;
- all studies retrieved from the searches are listed in tables at the end of the report (together with reasons justifying inclusion and exclusion).

It is helpful to set out the review protocol in a consistent format, which will aid peer reviewers. As an example, a protocol for a systematic review of randomized trials evaluating interventions in the teaching of spelling is shown in Box 4.1.

Box 4.1 Example of a review protocol

What is the title?
A systematic review and meta-analysis of randomized controlled trials evaluating interventions in the teaching of spelling.

What is the context and what are the conceptual issues?
There is a widespread consensus that spelling is a difficult skill. Indeed, some believe it to be more difficult than reading, because it requires the formation of an exact sequence of letters without any contextual clues (Fulk and Stormont-Spurgin 1995). Poor spelling skills are a widespread problem. Many children find spelling difficult, particularly those who experience learning difficulties (Fulk and Stormont-Spurgin 1995),

but including those who are high attainers in other areas (McClurg and Kasakow 1989). Many children continue to rely on phonetic strategies into their later years, and there is still controversy among teachers and researchers about the appropriate strategies for spelling instruction (McClurg and Kasakow 1989). Whilst spelling is often seen as a 'lower order' literacy skill based on memory, this view is simplistic: it is a highly complex ability.

Spelling acquisition, like other aspects of literacy, is developmental. Teachers use a variety of methods and instructional techniques to teach spelling skills, for example systematic study or structured study conditions, multisensory training, and spelling within the context of written composition. It is often unclear which is the most effective method of teaching spelling. Consequently most teachers use a variety of methods.

What is the policy context?
In the UK, The National Literacy Strategy for England and Wales gives detailed guidance for teaching and learning spelling (DfEE 1998). In the early years and throughout Key Stage 1 the emphasis is on phonological awareness, phonemic awareness and phonics teaching. At Key Stage 2 the emphasis for spelling is on individual self-correction strategies, independent spelling strategies (for example, phonics-based strategies, dictionaries and IT spell-checks), learning spelling conventions and rules (for example, patterns, prefixes and suffixes), practising spelling (using 'look, say, cover, write, check') and investigating the spelling of words (for example, word origins and derivations). Although the strategy is still highly contested, clearly it follows the now uncontested developmental model of spelling abilities. At Key Stage 1 the influence of international research synthesis on phonological awareness training, phonemic awareness training and systematic phonics teaching is evident (Troia 1999, Ehri, Nunes, Stahl, Willows 2001, Ehri, Nunes, Willows *et al.* 2001).

Has a scoping review been undertaken (if yes, what were the results)?
A scoping review was undertaken which found six relevant systematic reviews (Fulk and Stormont-Spurgin 1995, Troia

1999, MacArthur *et al.* 2001, Ehri, Nunes, Stahl, Willows 2001, Ehri, Nunes, Willows *et al.* 2001, Torgerson and Elbourne 2002). Two examined the effect of ICT on spelling acquisition (MacArthur *et al.* 2001, Torgerson and Elbourne 2002). The third review investigated phonological awareness training (Troia 1999); the fourth examined the effect of phonemic awareness training (Ehri, Nunes, Stahl, Willows 2001); the fifth systematically reviewed the experimental research on systematic phonics teaching versus non-systematic phonics teaching and the sixth reviewed published research on spelling interventions designed for pupils experiencing learning disabilities (Fulk and Stormont-Spurgin 1995).

What is the aim?
This aim of this review is to help fill the gap in the knowledge base of what does and does not work in the teaching of spelling.

What is the research question?
The research question for the scoping stage of the review is: which interventions or strategies are effective in the teaching of spelling for pupils aged between 5 and 16? The research question for the in-depth stage of the review is: which interventions or strategies are effective in the teaching of spelling for 'normally achieving' pupils aged between 7 and 14 (in the UK, Key Stage 2). The reason for limiting the review in this way at the in-depth stage is because of the existence of previous systematic reviews in early literacy development (Troia 1999, Ehri, Nunes, Stahl, Willows 2001, Ehri, Nunes, Willows *et al.* 2001) and spelling development with children experiencing learning disabilities (Fulk and Stormont-Spurgin 1995), and the lack of a systematic review focusing on interventions to improve the spelling abilities of 'normally achieving' children and young people aged between 7 and 14 (in the UK, Key Stage 2).

What is the search strategy?
The Educational Resources Information Center (ERIC); PsycINFO; and The Campbell Collaboration Social, Psychological, Educational Criminological Trials Register (C2 SPECTR) will be searched. All the searches will be for the

period 1980–2002. As the topic focus will be the teaching of the spelling of English, the search will be restricted to the English language research literature. In a previous review of ICT and spelling (Torgerson and Elbourne 2002) it was found that the key words of allocat* experiment* and random* was the most sensitive search strategy for the identification of trials. These key words combined with spell*, should be the most sensitive search strategy for this review.

What are the inclusion/exclusion criteria?
As the research question is looking at 'effectiveness', papers using rigorous methods to assess effectiveness will be required. In essence, this implies randomized controlled trials (RCTs). Therefore, only randomized controlled trials (or systematic reviews containing at least one RCT) will be eligible for inclusion. For a paper to be included it will have to be a trial comparing two or more methods or strategies for the teaching and/or learning of spelling in a school setting. RCTs will only be included if they are undertaken in English-speaking countries, and written in the English language. Trials will be included if they are published (or unpublished but in the public domain) in the years 1980–2003 and if all of the participants are aged between 5 and 16. To be included in the review a trial will have to report at least one spelling outcome measure.

How will the data be extracted and analysed?
Data about participants, interventions, outcomes and quality of the studies will be extracted from all the included papers, using a standard format. Included studies will be tabulated, and effect sizes for the main and secondary outcomes will be calculated. Standardized effect sizes will be estimated by dividing the mean differences between the groups by a pooled standard deviation using a commercially available software package.

The educational validity of pooling two or more of the trials in a meta-analysis will be estimated using expert opinion. If appropriate, trials with a similar contextual framework will be pooled in a meta-analysis. Statistical and educational sources of heterogeneity will be investigated and possible reasons described. Sensitivity analyses of the results will include country specific analysis as well as cross-national pooling,

high- versus low-quality trials. Potential publication bias will be explored using a funnel plot and also comparing effect sizes from published and unpublished reports.

How will the quality of studies be assessed?
The trials will be quality appraised using a checklist derived from the CONSORT guidelines (Altman 1996). These guidelines are used by the major medical journals when publishing randomized controlled trials and include, for example, the following internal validity criteria: are groups comparable at baseline? was 'intention to teach' used? were post-tests undertaken 'blind'? The external validity of included trials will also be examined.

References

Altman, D. G. (1996) 'Better reporting of randomised controlled trials: the CONSORT statement', *British Medical Journal*, 313, 570–1.

DfEE (1989) *The National Literacy Strategy. Framework for Teaching*: London: DfEE.

Ehri, L. C., Nunes, S. R., Stahl, S. A. and Willows, D. M. (2001) 'Systematic phonics instruction helps students learn to read: evidence from the national reading panel's meta-analysis', *Review of Educational Research*, 71, 393–447.

Ehri, L. C., Nunes, S. R. and Willows, D. M., Valeska Schuster, B., Yaghoub-Zadeh, Z. and Shanahan, T. (2001) 'Phonemic awareness instruction helps children learn to read: evidence from the National Reading Panel's meta-analysis', *Reading Research Quarterly*, 36 (3), July/August/September 250–87.

Fulk, B. M. and Stormont-Spurgin, M. (1995) 'Spelling interventions for students with disabilities: a review', *The Journal of Special Education*, 28 (4), 488–513.

MacArthur, C. A., Ferretti, R. P., Okolo, C. M. and Cavalier, A. R. (2001) 'Technology applications for students with literacy problems: a critical review', *The Elementary School Journal*, 101 (3), 273–301.

McClurg, P. A. and Kasakow, N. (1989) 'Wordprocessors, spelling checkers, and drill and practice programs: effective tools for spelling instruction?', *Journal of Educational Computing Research*, 5, 187–98.

Torgerson, C. J. and Elbourne, D. (2002) 'A systematic review and meta-analysis of the effectiveness of information and communication technology (ICT) on the teaching of spelling', *Journal of Research in Reading*, 35 (2), 129–43.
Troia, G. A. (1999) 'Phonological awareness intervention research: a critical review of the experimental methodology', *Reading Research Quarterly*, 34, 28–52.

The review protocol is an important first step when undertaking a systematic review. It helps to focus and structure the review; it limits the scope for bias occurring in the review and enables an independent third party to critically appraise the finished review in relation to the initial proposal.

Literature search

The main thrust of the search should be in the electronic databases, as being the most efficient method of retrieval (NHS Centre for Reviews and Dissemination 2001). This is mainly due to the technological explosion in the last ten to fifteen years and the subsequent availability of educational and other relevant electronic databases that can be systematically searched. However, there are many other sources of educational studies: key journals; the bibliographies of systematic and other reviews; websites; personal contact with content specialists. Whilst using all of the above methods for retrieval may make the search more exhaustive, some of the methods can increase the risk of potential selection bias. This problem is discussed below.

The three methods that are least liable to selection bias are searching electronic databases, hand searching of key journals and searching the bibliographies of previous systematic reviews, because all of these methods employ a

'systematic' approach. Therefore this book will confine itself to an overview of how randomized trials might be identified using these three methods.

Ideally before undertaking a search of electronic databases it is extremely helpful to enlist the aid of a librarian or information specialist who will know how to access the most relevant databases and will be able to advise on the appropriate search strategy for each database.

Electronic searching

A preliminary search of the main electronic databases often takes place to aid the development of the review protocol (rapid scope). Health care reviewers, who are focusing only on randomized controlled trials, are in the fortunate position of being able to search a single database (the Cochrane Library) that contains all RCTs relating to health. Presently, educational and other non-health care reviewers are less fortunate. In future the Campbell Library may rival the Cochrane Library in its completeness but this is not the case at present. In addition, the EPPI-Centre is developing an electronic database of reviews in educational research. At present, however, the educational researcher must search across a number of electronic databases to ensure that the majority of RCTs within a given area are identified.

For any search strategy there is a trade-off between 'sensitivity' and 'specificity'. A search strategy that is extremely sensitive is likely to include as many of the relevant studies as possible that are present on a database. Clearly the most sensitive strategy is a search of the entire database. Specificity relates to the concept of 'homing' in on the most relevant papers. Any search strategy that does not involve screening the entire database introduces specificity. The trade-off between sensitivity and specificity begins as soon as terms are introduced that start to exclude studies. It is

absolutely essential to introduce search terms to reduce the huge volume of literature within the educational databases. However, even the most specific search strategy will reveal many papers that could be RCTs and the only way to ascertain whether they are or not is to retrieve them and check in the methods section what the researchers actually did.

Designing the search strategy

The use of the 'wildcard' characters * and $ is important for devising appropriate research terms. Studies using random allocation may keyword this in numerous ways in the title or abstract. For example, a study might state: 'we undertook a randomized trial' (with either 's' or 'z'); or 'we randomly placed'; or 'we performed random allocation'; or 'groups were formed by using random number tables'. Rather than trying all the different methods of using the word 'random' a search term with random* will identify all the words with the stem random. Educational and psychological databases actually do not often use the term random* when describing randomized controlled trials. A widely used word is 'experiment' or 'experimental' with the term experiment* being appropriate for the search. Many researchers in the educational field seem to be rather hesitant about describing their study as a trial or an experiment and sometimes they will state in the abstract that participants were 'allocated into two groups'. Often in such studies true randomization has occurred but this is only apparent when the full paper is retrieved and the methods section scrutinized. Therefore, the term allocat* should also be used to identify studies which contain 'allocate', 'allocation' and 'allocated' in their title and abstract. A search term using the phrase 'allocat* or experiment* or random*' should include most of the experimental studies from an educational database.

The Campbell Collaboration has developed a search strategy that helps identify controlled trials within the educational databases (see Petrosino *et al.* 2000, Appendix 1).

Where are the randomized controlled trials in educational research?

There are a number of electronic databases that can be used to search for randomized controlled trials. These include PsycINFO (a database of psychological literature); ERIC (Educational Resources Information Center); BEI (British Educational Index); C2-SPECTR (Campbell Collaboration Social, Psychological, Educational and Criminal Trials Register) and SSCI (Social Sciences Citation Index). In addition, the database of 'grey' literature (i.e. unpublished literature in the public domain, e.g. reports, theses) can be searched (System for Information on Grey Literature in Europe: SIGLE), although ERIC is also a good source for this kind of literature.

Each of these databases has its strengths and weaknesses. With respect to searching for randomized controlled trials PsycINFO and ERIC are, at present, the richest sources of controlled studies for most educational reviews. However, as stated previously, the establishment of the Campbell Collaboration may change this situation in the future.

As an example, consider the findings of two recent systematic reviews: one reviewing research on the effectiveness of information and communication technology on literacy (Torgerson and Zhu 2003), and a systematic review of interventions to increase adult literacy and numeracy (Torgerson, Porthouse, Brooks 2003). These two reviews contained a total of 51 RCTs: 42 RCTs in the review of ICT and literacy and nine in the adult literacy and numeracy review. Table 4.1 shows the origin, by method of retrieval (electronic database or hand search) of all these RCTs.

Table 4.1 Origin of identified RCTs

Electronic database/hand search	Number of RCTs
PsycINFO	29
ERIC	16
SSCI	2
BEI	2
Hand search	2

Fifty-seven per cent of the RCTs were retrieved from PsycINFO with a further 31 per cent identified from ERIC. Together these two databases contained 88 per cent of all the RCTs identified for these two reviews. The other databases produced only a few additional references, as did hand searching of key journals. Nevertheless, if the search had been confined to PsycINFO and ERIC this would have led to the loss of more than a tenth of the relevant trials.

Hand searching

Because trials in educational research tend not to be specifically 'tagged' in educational databases, unlike their health equivalents, hand searching may be helpful in identifying some papers. The choice of journals to hand search should be based on which journals are most likely to yield the relevant trials. Sometimes the electronic searches can aid this decision by indicating which journals are the sources for most of the trials. Additionally, some journals, especially the newer ones, have not yet been indexed on the electronic databases and therefore the only way that relevant articles can be retrieved from these journals is through hand searching. Searching obscure journals can be challenging. If university libraries do not hold them, sometimes the only option is to spend time at a national library (e.g. in the UK the British Library in London or its

lending division at Boston Spa in Yorkshire; in the US the Smithsonian Institute in Washington).

Searching reference lists of systematic reviews

Previous systematic reviews can be helpful in identifying relevant papers. Reviews that address similar research questions or include questions that overlap with the systematic review being undertaken may contain relevant references.

Sources of search bias

The searching process can be prone to bias. Many reviewers, for example, use 'personal knowledge' or 'personal contacts' to identify relevant studies. In theory this can result in bias if the same studies are not identified by electronic or hand searching. The problem with personal knowledge is that studies favouring a particular viewpoint are more likely to have been noted and retained whilst other studies not favouring this perspective may have been discarded. On the other hand, personal knowledge may reveal key unpublished trials, which if not included, could result in the review coming to erroneous conclusions. Therefore, it is not immediately clear whether such references should be included or not. A useful compromise is to include them but examine the results of the review when they are excluded (i.e. by undertaking a sensitivity analysis).

Some reviewers also search the bibliographies of non-systematic reviews as a means of identifying relevant studies. The problem with this is that non-systematic reviews may include a biased sample of the literature and this bias could then be introduced into the systematic review.

Screening

Once the search strategy has been determined, and potentially relevant titles and abstracts identified, the next step is to filter out the irrelevant papers and screen in possibly relevant articles. The first step in this process is to 'de-duplicate' the references by importing them into a reference management software package (e.g. EndNote; Pro-Cite; Reference Manager). Establishing a database of references enables a record to be kept of every step of the review, which if necessary can be rerun by a third party.

Potentially relevant studies are identified from titles and abstracts (first stage screening). Irrelevant papers are filtered out and potentially relevant papers are sent for. These are then read and identified as either being relevant or not (second stage screening). These processes are, ideally, undertaken by two independent researchers to ensure that only a minimal number of relevant studies are 'missed'. If reviewers agree on references these can be either discarded or retrieved. If there is disagreement the references can be examined by both reviewers and discussed, at which stage they can either be included or rejected. Double screening, however, is resource intensive. An alternative strategy is for the database to be screened by one reviewer only. A second reviewer can then screen a random sample of the database (e.g. 10 per cent). This random sample of citations can be used to measure the inter-rater reliability assessment of the agreement between the reviewers. The statistic, Cohen's Kappa, can then be calculated to measure how well the two raters agree. This takes into account the agreement that would have occurred by chance. The values range between +1 (perfect agreement) and 0. A value of 0 indicates that the observed agreement could have occurred by chance. This process describes how well the decisions could be reproduced.

Table 4.2 A worked example of inter-rater reliability assessment using Cohen's Kappa

	Screener B		
Screener A	Include	Exclude	Total
Include	85	60	145
Exclude	59	4320	4379
Total	144	4380	4524

If the level of agreement, as measured by Cohen's Kappa, is high (for example 0.90) then the results of the single reviewer can be relied upon. However, if agreement is low (for example 0.25), then it would be necessary for the second person to double screen all of the citations.

The process of calculating Cohen's Kappa is undertaken with reference to Table 4.2 as follows. An electronic screening strategy identified 4524 potential articles and two reviewers screened this database. To find out how well they agree the exact number of agreements is calculated, which in Table 4.2 is 85 + 4320 = 4405. From a total of 4524 articles this is 0.97 or 97 per cent (i.e. 4405/4524). This overall agreement figure, however, takes no account of the fact that some of the agreement will occur by chance. The next step is to take this chance effect into account.

Include	144 × 145/4524 =	4.62
Exclude	4379 × 4380/4524 =	4239.61
Total		4244.22

The number of agreements that is expected by chance, therefore, is 4244.22, which as a proportion of the total articles is 4244.22/4524 = 0.938. The maximum agreement is 1.0, therefore, we can calculate the inter-rater reliability agreement as:

$$\frac{0.97 - 0.938}{1.00 - 0.938}$$

which results in a Kappa value of 0.52 (moderate).

Table 4.3 presents the results of screening a search for relevant randomized controlled trials, controlled trials and reviews for a systematic review of interventions to increase adult literacy and/or numeracy (Torgerson, Porthouse, Brooks 2003). The number of relevant studies can be extremely small: about one-third of 1 per cent for RCTs in the case of this review. In this instance the database of 'Criminal Justice Abstracts' was searched because, although it is not an educational database, it was known that controlled trials had been undertaken in prison settings. Five CTs and reviews were retrieved from this database although none of these was a randomized controlled trial.

Table 4.4 is a 'scoping' or 'mapping' table, which shows the numbers of studies that were retrieved after screening of the electronic searches, and the reasons why only twelve of

Table 4.3 Origin of all included studies

	Found	Included RCTs, CTs and systematic and other reviews	Number of RCTs
ERIC	2628	40	9 (0.34%)
PsycINFO	971	3	2 (0.21%)
CJA	736	5	0
SSCI	15	2	2 (13%)
C2-SPECTR	8	1	0
SIGLE	172	1	0
Website	11	2	1 (9%)
Bibliography search	13	4	1 (8%)
Contact	1	1	0
Total	4555	59	15 (0.33%)

Table 4.4 Mapping of relevant RCTs, CTs and reviews

RCTs	12 papers (containing 9 trials)
RCTs (no results)	3
CTs	34
Reviews	10
Total	59

the 59 papers were deemed relevant for the in-depth review. Because the requirements of this review were to identify primarily RCTs but also controlled trials (CTs) both are listed in the mapping table, along with relevant review articles. However, because the inclusion criteria for this review specified including only RCTs in the in-depth review only nine RCTs (reported in twelve papers) were data extracted and quality appraised. The other papers (controlled trials and reviews) were used to provide conceptual background information.

In Tables 4.5 to 4.7 the process of screening and mapping the literature within a given field, in this case a systematic

Table 4.5 Screening and mapping of the literature in a systematic review on spelling

	Found and screened	Ex-first stage*	Sent for	Ex-second stage**	Not received	Included in map***	Included in in-depth
ERIC	178	131	47	31	1	15	3
PsycINFO	311	265	46	27	1	18	7
Expert contact	6	1	5	0	0	5	0
Bibliographic searches	3	0	3	0	1	2	1
Total	498	397	101	58	3	40	11

* Screening on the basis of titles and abstracts; all 'includes' sent for

** Screening on the basis of full papers received

*** Description of the research in the field

Table 4.6 Reasons for first and second stage exclusions

	First stage exclusions	Second stage exclusions
Not spelling	106	0
Not spelling intervention	27	8
Not English	52	5
Not trial	140	8
Age 1 (not 5–16)***	14	1
Not randomized controlled trial (or systematic review containing at least one RCT)	58	32
Not spelling only	0	4
Total	397	58

*** For scoping stage RCTs evaluating interventions in pupils aged between 5 and 16 were included

Table 4.7 Third stage exclusions (29 papers in the 'map' of the research, but excluded from the in-depth review)

Learning disabilities	10
Age 2 (not 7–14)**	12*
Reading disabilities	2
Reading	1
Systematic review	4
Total	29

* Two of these papers contain the same study

** For in-depth review only RCTs evaluating interventions in pupils aged between 7 and 14 (Key Stage 2) were included. This was because the in-depth review excluded early literacy interventions (phonological awareness and phonemic awareness training)

review of spelling interventions, is described. Table 4.5 shows the origin of the studies that were identified in the original electronic search. The original number of studies, 498, is reduced through a series of predefined exclusion criteria until only eleven relevant studies are left for the in-depth systematic review. Tables 4.5 and 4.6 describe the

reasons for excluding studies. Most studies were excluded at the first stage because they were not studies about spelling or because they were not trials. It is quite usual to exclude a large percentage of studies at the first stage because they are outside the scope of the review. Most studies were excluded at the second stage of screening because they were not randomized controlled trials.

All of the papers in the 'map' were randomized controlled trials. However, for the in-depth review only RCTs evaluating interventions to improve spelling in 'normally achieving' children aged 7–14 were included. The remaining eleven papers in the in-depth review fulfilled these criteria.

Data extraction

Once screening has been completed and potentially relevant studies identified, the relevant data for the review are extracted onto a standardized data extraction form. Box 4.2 gives an outline of a minimum data extraction sheet that can be used to extract the key data from randomized trials.

Box 4.2 Example of data extraction sheet

Author:	
Year:	
Country:	Country where the research was carried out.
Publication type:	Journal article; book chapter; unpublished dissertation; report.
Reference:	Full reference including title of journal, volume, page numbers.
Source:	Where the reference was identified (e.g., PsycINFO, hand search).
Setting:	Setting where study was carried out (e.g., elementary school).
Objective:	Objective of the study as stated by the authors.

Outcome measures:	All outcome measures as stated by the authors in the methods section.
Design:	Type of RCT (e.g., cluster or individual or cross-over).
Participants:	Detailed description of participants involved in study (e.g., age, gender, ethnicity, socio-economic factors, learner characteristics).
Intervention:	Detailed description of the intervention.
Control:	Detailed description of control treatment.
Results (as reported by authors):	All results including those in narrative and in tables.
Effect size (as reported by authors):	If authors report effect size(s) include.
Effect size (as calculated by reviewers):	If authors do not include effect size reviewers need to calculate.
Comments:	Details about study quality (e.g., attrition rate); if the RCT is reported elsewhere give reference.

Data extraction sheets should be piloted on several trials and amended if necessary before extracting data on all the trials. As part of its methodological work in research synthesis the EPPI-Centre has developed detailed data extraction guidelines and tools for use with the DfES-funded EPPI-Centre review groups (see http://eppi.ioe.ac.uk/EPPIWeb/home.aspx). It has been suggested that data extraction should be undertaken by the reviewer 'blind' to the article's authors and the journal to reduce the risk of bias (Egger and Davey-Smith 2001). The reasoning behind this is that reviewers might be more favourable in their judgement towards a paper that they know has been published in a prestigious journal or has originated from a highly respected research group. However, masking reviewers to the identity of papers is extremely time-

consuming as all papers need to be photocopied with the authors and identifying features of the journal removed by an independent person. However, a randomized trial comparing blinded with unblinded data extraction found no significant difference in results between the approaches (Berlin 1997).

Box 4.3 is an example of a completed data extraction sheet from a systematic review of randomized controlled trials evaluating interventions in adult literacy and numeracy (Torgerson, Porthouse, Brooks 2003). Data were extracted using the data extraction sheet in Box 4.2. The paper was double data extracted by two reviewers who then discussed the data extraction and resolved any differences. The sheet summarizes the key aspects of the trial, for example bibliographic details and information about the aims of the study, the intervention and the outcomes measured, characteristics of the participants and a summary of the results of the trial.

Box 4.3 Example of completed data extraction sheet

Author:	Batchelder and Rachal
Year:	2000
Country:	USA
Publication type:	Journal article
Setting:	Maximum security prison
Objective:	To examine the efficacy of using computer-assisted instruction (CAI) with inmates participating in a prison education programme compared with inmates participating in a traditional instruction programme using an experimental design.
Study topic:	Literacy and numeracy CAI Incarcerated population
Outcome measures:	Comprehensive Adult Student Assessment System (CASAS) maths and reading post-tests.

Design: RCT (individual), digit table.

Participants: n = 75 male inmates in maximum security prison.
Two ethnic groups: African-American inmates (n = 56) and Caucasians (n = 15).

Intervention: I: Participants received GED instructional material for 1 hour per day on computers for a total of 80 hours over a 4-week period in mathematics or language. Also traditional instruction for 3 hours per day in English, maths, history and science.

Control: Participants received traditional instruction in English, maths, history and science for 4 hours per day for a total of 80 hours over a 4-week period.

Results: (as reported by authors): Achievement scores of inmates in the intervention group were not significantly higher than those in the control group.

Group	Mean	SD	n
CASAS maths post-test			
Group 1: Experimental	221.9	12.3	36
Group 2: Control	217.0	17.9	35
CASAS reading post-test			
Group 1: Experimental	227.4	13.5	36
Group 2: Control	223.4	17.5	35

Effect size: CASAS maths: No significant difference
Unadjusted effect size = 0.16
CASA reading: No significant difference
Unadjusted effect size = 0.26

Effect size (as calculated by reviewers): No difference between I and C

Comments: Study also reported as: Batchelder 2000
Attrition: n = 4

A key issue with respect to data extraction is the reviewer decision about whether a study is actually a randomized trial or not. Slavin (1986) has noted that some meta-analyses of 'randomized controlled trials' in education included non-randomized controlled trials. The reviewers were unclear of the difference. In a randomized controlled trial participants are *randomly assigned* to their instructional group (intervention or control). Unfortunately, randomization as a procedure for allocating individuals or clusters to an intervention or control group is not always well understood by educational researchers (Fitz-Gibbon 2000) and often not clearly described.

Phrases that are often used to describe random allocation in educational trials include:

> 'Children were paired on the basis of gender and age and allocated *randomly* using random number tables or coin toss to the treatment'
>
> 'Using random number tables or coin toss children were assigned to their groups'
>
> 'Using restricted or stratified or blocked allocation schedule participants were assigned to their groups'
>
> 'Intact classrooms or schools were matched on class size and a member of each pair was randomly assigned to the intervention'

In contrast the following do *not* describe random allocation:

> 'We took a random sample of children from schools that were not implementing the curriculum and compared them with a random sample of children in the intervention schools'
>
> 'After the schools/children/students had been randomly assigned we asked teachers to identify, for post-test, those children who they felt had benefited most from the intervention'

> 'Two schools were chosen to take part and one school was randomly allocated to receive the new curriculum'

The last statement may appear to be a 'cluster randomized trial' but it is not. Randomizing schools is a perfectly legitimate method of performing a randomized controlled trial; however, if there is only one cluster in each arm of the trial this cannot control for school effects, whether random allocation was used or not. It is recommended that two-armed cluster trials should have at least eight clusters (i.e. four in each arm) and preferably more to allow randomization to balance out any school level confounders (Ukoumunne *et al.* 1998).

Even knowing whether to describe a trial as being randomized or not can sometimes be difficult to ascertain. For example, it is not clear from the published report of a trial evaluating phonological awareness training by Hatcher and colleagues (1994) whether or not it is a randomized controlled trial. Therefore, if one relied solely on the published report such a study would probably be classified as a controlled study, not the more rigorous randomized trial. However, the children in that study were actually allocated to their treatment groups in a randomized fashion (Hatcher, personal communication, 2001). In some reviews, therefore, it may be necessary to contact the authors for more details about their studies to facilitate both data extraction and quality appraisal.

Sometimes studies claim to have produced matched pairs of children or students and then randomly allocated one member of a pair to the intervention. This process should produce exactly equal numbers in each group. The presence of uneven numbers in intervention and control groups when using a matched-pair design gives cause for concern about the quality of the study. In contrast, when small studies use simple allocation it is perfectly possible to have exactly equal numbers but this is unlikely. Nevertheless many small

studies using 'simple' randomization or random number tables have suspiciously good numerical equivalence. Numerical balance, in small trials, is only likely if some form of stratified allocation mechanism is used. In contrast, large trials, whilst unlikely to have exact numerical balance, should have an approximate 50:50 split. An interesting example of inconsistent allocation occurred in a study of an adult education programme in six counties in California. Altogether 20,000 participants were randomized (by simple allocation methods). There were approximately equal numbers of participants in only one county; in the remaining five counties the percentage allocated to the experimental group ranged from 68 per cent to 86 per cent (Martinson and Friedlander 1994, cited in Torgerson, Porthouse, Brooks 2003). This disparity in group-size was never satisfactorily explained.

Summary

- The protocol is an *a priori* statement of the research question, aims and methods of the review. It includes the procedures for searching and screening, data extraction, quality appraisal and synthesis.
- The literature search should focus on the electronic databases, but may include other methods of retrieval, for example hand searching of key journals, searching bibliographies of other reviews, personal contacts.
- Ideally, screening, data extraction and quality appraisal should be undertaken by two researchers, working independently.

5

Quality Appraisal

The main reason for undertaking a randomized controlled trial is to obtain evidence with a high degree of internal validity. Although RCTs are widely regarded as the 'gold standard' of effectiveness research, clearly their results are more reliable when the trials are of high quality. Over the last decade trial methodologists working in the health field have developed a set of guidelines that trialists should adhere to if they wish to report a good-quality trial – these have been published as the Consolidated Standards for Reporting Trials (CONSORT statement) (Altman 1996). The motivation for CONSORT was the poor quality of so many of the RCTs that have been published in the health care field, which may misinform policy. Many major medical journals now insist that reports of RCTs conform to the CONSORT guidelines (Altman 1996).

Low-quality trials have also been undertaken and published in the field of educational research. In a systematic review of the effects of information and communication technology (ICT) on the teaching and learning of spelling it was noted that the quality of trials included in that review was generally low (Torgerson and Elbourne 2002). At present in the field of educational research there is no equivalent of the CONSORT statement. However, educational researchers have long recognized the need to 'quality-appraise' RCTs in education (Slavin 1986, Troia 1999, Torgerson and Elbourne 2002).

The issue of trial quality has increased in importance in

the field of health care research. Methodological reviews have described a relatively high prevalence of poor-quality trials, which can mislead health care practice and policy (Schulz *et al.* 1995, Kjaergard *et al.* 2001). Indeed, recently a large methodological analysis sought to explain the puzzling phenomenon of larger trials yielding smaller effect sizes, on average, than smaller trials, even when they are attempting to address the same question. Kjaergard and colleagues (2001) examined the quality of large and small trials and found that large trials tended to be of better quality than small studies. After they had taken quality of trial methodology into account the difference in observed effect sizes between large and small studies disappeared. This indicates, therefore, that poor-quality studies, rather than small trials, could be a source of bias when included in a meta-analysis. This problem is likely to affect educational trials as well as health care studies. For example, Lipsey and Wilson (1993) noted that educational and psychological trials with sample sizes of more than 100 yielded smaller average effect sizes compared with smaller trials. More recently, in a systematic review of phonemic awareness training Ehri, Nunes, Stahl, Willows (2001) found effect size was inversely related to sample size in reading and spelling outcomes. It is possible that size may be a marker for poor trial quality in education just as it is in health care.

Educational researchers are aware of the potential problems of poor-quality trials and many have produced sets of quality criteria in order to classify studies as being rigorous or not. For example the EPPI-Centre has developed detailed guidelines and tools for quality appraisal of randomized controlled trials and all other study types. Table 5.1 contains an example of quality criteria that were developed to assess the quality of controlled trials in phonological awareness training (Troia 1999). As well as listing the various criteria, often a 'scoring' system is used so that a summary score can be given to a trial.

Table 5.1 Troia's study quality criteria relating to internal validity

Validity criteria	Weighting
Random assignment	3
Control group received alternative intervention to control for Hawthorne effect	3
Exposure to similar materials for control group	1
Counterbalancing of teachers	2
Treatment explicitly described	2
Criterion-based intervention	1
Equivalent instructional time	3
Equivalent mortality rates	1

Measurement of study quality, however, is not necessarily an objective exercise. The use of any quality score can be fraught with difficulty. For example, Juni and colleagues quality appraised seventeen health care trials, from a meta-analysis, with 25 different quality scales (Juni *et al.* 1999). They found that, for twelve scales, the effect sizes were the same when trials were rated as high or low quality. However, for six scales, high-quality trials showed little or no benefit of treatment compared with low-quality studies, whilst the remaining seven scales showed the opposite. Thus, quality assurance scales can give very different results depending on the items included and the weights given to individual items.

If quality criteria use a system of 'weighting' or 'adding up' there is a risk of classifying a trial as being of 'good' quality simply because it performs well on many of the criteria. However, if the trial has a fatal flaw in one of the most important aspects of trial design, the results of the trial may be unreliable. On some scales studies can score highly if they are well reported rather than well conducted (Juni *et al.*

2001). Juni and colleagues (2001) explain how a trial can be defined as being of 'high' quality on one widely used scale in assessing the quality of health care trials even if the authors report that they did *not* use random allocation because the scale emphasizes reporting rather than actual performance.

Many aspects of study or trial design can affect the outcome of a study. The most important design criteria relate to its internal validity. If a study is not internally valid, then the observed effect sizes from a study may be incorrect. Clearly the study design with the greatest internal validity is the experimental method using randomization to assemble comparable groups (Cook and Campbell 1979). Despite randomization, however, forms of selection bias can be introduced during the trial (see Torgerson, and Torgerson, 2003b for a full discussion). If researchers subvert the allocation schedule this can introduce a source of bias. This phenomenon has been documented both in health care trials (Schulz *et al.* 1995) and criminal justice studies (Boruch 1997). Ideally, trial allocation should be undertaken by an independent person, as this will reduce the risk of the allocation being subverted. A symptom of a problem with randomization is if the 'baseline' variables differ between the groups (i.e. there is baseline imbalance).

Once randomization has occurred bias can still be introduced if outcomes are not measured blindly at post-test. If the researchers or assessors are aware of the allocated group they may, consciously or unconsciously, give higher marks to those students in one group. Outcome assessment should be undertaken by someone who is 'masked' or 'blinded' to group assignment (Cook and Campbell 1979).

Two other important aspects of trial design include attrition and intention to teach. Attrition, often referred to as 'mortality' in educational papers (Troia 1999), is when participants drop out of the study between randomization and post-test. If the drop-out rate is either high or unequal between the groups then this can introduce selection bias.

Those who drop out from one group may be different from those who remain in the comparison group. Another analytical problem occurs when not all participants are included in the final analysis. Some researchers undertake 'active treatment' analysis, that is only analysing participants if they receive the intervention to which they were allocated. The most rigorous way to analyse the data is to undertake 'intention to teach' analysis. This is where all participants are analysed in the groups into which they were originally allocated. This may be difficult to achieve in practice as some participants usually 'drop out' and therefore they cannot be included in post-tests.

In summary, to ensure a robust and valid trial one should look for concealed randomization; similar attrition rates; no baseline imbalance; blinded or masked follow-up. Table 5.2 contains a modified version of the CONSORT criteria, widely used in health care, which can be used to describe the quality of trials identified in education.

The most important aspects of quality relate to the internal validity of the trial and these are highlighted in italics in Table 5.2. There are, however, other important aspects of trial quality that are included in the CONSORT quality check. Three of these relate to the issue of sample size or the possibility of a Type II error. A Type II error occurs when there is a 'true' difference between groups, but the sample size is insufficient to demonstrate this difference as being statistically significant. The larger the study, the less likely it is to suffer a Type II error.

Sample size

In the field of education, as in health care, most effective experimental innovations yield small to moderate positive effects (Kulik and Kulik 1989, Lipsey and Wilson 1993). Therefore, researchers seeking statistical significance must

Table 5.2 Modified CONSORT quality criteria

Was the study population adequately described? (i.e. were the important characteristics of the randomized participants described, e.g. age, gender?)
Was the minimum important difference described? (i.e. was the smallest educationally important effect size described?)
Was the target sample size adequately determined?
Was intention to treat analysis used? (i.e. were all participants who were randomized included in the follow-up and analysis?)
Was the unit of randomization described (i.e. individual participants or groups of participants)?
Were the participants allocated using random number tables, coin flip, computer generation?
Was the randomization process concealed from the investigators? (i.e. were the researchers who were recruiting participants to the trial blind to the participant's allocation until after that participant had been included in the trial?)
Were follow-up measures administered blind? (i.e. were the researchers who administered the outcome measures blind to treatment allocation?)
Was estimated effect on primary and secondary outcome measures stated?
Was precision of effect size estimated (confidence intervals)?
Were summary data presented in sufficient detail to permit alternative analyses or replication?
Was the discussion of the study findings consistent with the data?

use large sample sizes. The probability of an 'educationally significant' difference being also statistically significant is partly a function of sample size. Small sample sizes can miss important differences between the treatment groups. Importantly for systematic reviewers, small sample sizes often lead to null or non-significant *negative* results, which

can lead to the study not being published. Such trials will be excluded from any review and only positive, statistically significant, trials will be included. This will lead to an over-optimistic assessment of the benefit of a given intervention.

In order to ascertain whether or not a study is large enough, an educationally significant difference needs to be calculated. In their review of educational and psychological experiments, Lipsey and Wilson (1993) found that for effective interventions the effect sizes ranged, on average, from about 0.25 to 0.50. If it is assumed that, as a minimum, a trial ought to be large enough to detect at least half an effect size then, statistically it can be demonstrated that to have an 80 per cent chance of detecting half a standard deviation difference between two groups with a significance level of 5 per cent, a trial requires 63 children in each group (i.e. 126 in total). Trials smaller than this run a high risk of missing an important difference in outcome between the experimental and control groups. Indeed, even a sample size of several hundred would be too small to detect the benefit observed in the Tennessee experiment of class sizes. To observe the modest benefit of smaller class sizes would require several thousand children. Whether this small benefit is educationally 'significant' or 'worthwhile' is a matter for teachers, parents and policy-makers to debate.

Confidence intervals

The point estimate of an effect from any trial is bounded by uncertainty. For instance, a large effect can be statistically insignificant because the sample size is too small. One way of representing the boundaries of uncertainty around an estimate of effect is to use confidence intervals (usually 95%). The confidence interval represents (given the constraints of the sample size) where 95 per cent of the results would lie, if the experiment were repeated 100 times. Confidence intervals are important because they show the

uncertainty that surrounds the point estimate of effect. For example, Weiner (1994) showed an effect of phonemic awareness training that was not statistically significant with an effect size of about 0.3 among 30 pupils, which is considered a reasonable effect size in educational research. The upper confidence limit included an even larger effect size of about 1.0, indicating that the trial was too small to exclude a very large difference in effect. There was a real danger, therefore, that this trial experienced a Type II error: that is erroneously concluding there was no effect when in fact there was one.

In Table 5.3 the CONSORT quality criteria are applied to a sample of RCTs from a systematic review of interventions in adult education. This table is fairly representative of the reporting quality of educational trials. No trial report, for example, outlines the reasoning behind sample size calculation or reports confidence intervals. For internal validity criteria no trial reports whether the randomization process was undertaken independently and few trials report blinded outcome assessment and intention to teach analysis.

Note in this table (Table 5.3) for quality assurance, quality criteria are not given a weight or simply added up. It may be advisable for the reviewer to make a judgement relating to individual trials as to whether a given quality criterion that has not been fulfilled represents a 'fatal flaw' and undermines that study's results. In other words, undertaking a systematic review is not a mechanistic exercise: it requires skill and experience to interpret the results.

Summary

- The most important aspects of trial design relate to internal validity. If a RCT is not internally valid, the observed effect sizes may be incorrect.
- The quality of randomized trials included in systematic reviews should be assessed on their internal validity.
- Various sets of quality criteria to appraise RCTs have been developed.

Table 5.3 Example of the 'CONSORT' quality checklist applied to RCTs

	Batchelder and Rachal 2000	Bean and Wilson 1989	Cheek and Lindsey 1994	Martinson and Friedlander 1994	McKane and Greene 1996	Nicol and Anderson 2000	Rich and Shepherd 1993	Shrum 1985	St Pierre *et al.* 1993
Was the study population adequately described? (i.e. were the important characteristics of the randomized adults described, e.g. age, gender?)	Y	Y	Y	N	N	Y	Y	Y	N
Was the minimum important difference described? (i.e. was the smallest educationally important effect size described?)	N/S	N/S	N/S	N/S	N/S	N/S	N/S	N/S	N/S
Was the target sample size adequately determined?	N/S	N/S	N/S	N/S	N/S	N/S	N/S	N/S	N/S
Was intention to treat analysis used? (i.e. were all adults who were randomized included in the follow-up and analysis?)	N	N	N	N	N	N/S	Y	N/S	N

	Batchelder and Rachal 2000	Bean and Wilson 1989	Cheek and Lindsey 1994	Martinson and Friedlander 1994	McKane and Greene 1996	Nicol and Anderson 2000	Rich and Shepherd 1993	Shrum 1985	St Pierre *et al.* 1993
Was the unit of randomization described (i.e. individual adults or groups of adults)?	Y ind	Y Ind	Y ind	Y Ind	Y ind	Y ind	Y ind	Y ind	Y ind
Were the participants allocated using random number tables, coin flip, computer generation?	Y	N/S	N/S	N/S	N/S	N/S	N/S	Y	N/S
Was the randomization process concealed from the investigators? (i.e. were the researchers who were recruiting adults to the trial blind to the adult's allocation until after that adult had been included in the trial?)	N/S	N/S	N/S	N/S	N/S	N/S	N/S	N/S	U
Were follow-up measures administered blind? (i.e. were the researchers who administered the outcome measures blind to treatment allocation?)	N/S	N/S	N/S	N/S	N/S	N/S	Y	N/S	Y
Was estimated effect on primary and secondary outcome measures stated?	Y	Y	Y	U	Y	Y	Y	Y	Y

	Batchelder and Rachal 2000	Bean and Wilson 1989	Cheek and Lindsey 1994	Martinson and Friedlander 1994	McKane and Greene 1996	Nicol and Anderson 2000	Rich and Shepherd 1993	Shrum 1985	St Pierre *et al.* 1993
Was precision of effect size estimated (confidence intervals)?	N/S	N	N	N/S	N/S	N	N/S	N/S	N
Were summary data presented in sufficient detail to permit alternative analyses or replication?	Y	Y	Y	U	Y	Y	Y	N	Y
Was the discussion of the study findings consistent with the data?	Y	N	Y	N	Y	N	Y	Y	Y

N/S = not stated; U = unclear

Source: Torgerson *et al.* (2003) 'A systematic review and meta-analysis of randomised controlled trials evaluating interventions in adult literacy and numeracy', *Journal of Research in Reading*, 26 (3), 2003.

6

Publication Bias

A systematic review of randomized controlled trials provides an unbiased estimate of the effect of an intervention if one of two conditions is fulfilled: firstly, if *all* the relevant trials are included in the review; or, secondly, if a random *sample* of all the trials ever undertaken is included in the review. In many reviews it is unlikely that the first condition will be fulfilled. If the literature is particularly large then it is likely that some trials will be overlooked. As long as the trials that are missed do not depart in any significant or systematic way from the trials that are included, then the estimate of the intervention effect will not be biased. In a meta-analysis, a statistically non-significant estimate of effect could be due to missing trials. This is because the greater the number of trials included in the meta-analysis the more precise will be the estimate of effect. Trials 'missing at random' will not, on average, alter the direction of the effect size, but their non-inclusion will reduce precision.

A more serious problem occurs, however, when trials that are not included in the review are missing because they are unpublished, and have characteristics that make them different from published studies. Unpublished studies tend to demonstrate negative or null effects. Therefore, a systematic review will tend to retrieve a sample of trials that are positive, which will give an inflated estimate of any effect of the intervention. If the search strategy for a systematic review includes searching bibliographies of non-systematic reviews these will tend to cite the positive trials more often than the

negative studies. It is perfectly possible, therefore, by excluding trials 'not missing at random' to either overestimate the effectiveness of an intervention or, more seriously, to reverse the direction of effect. This may result in a review concluding that a harmful intervention is actually beneficial.

Historically, some journals have not published trials that do not produce 'significant' or different findings. Some journal editors and referees, therefore, will reject as being 'uninteresting' a trial that reports no difference between an intervention and control group. Rejection of a paper by journals and referees may be accompanied by a 'scientific rationale' justifying the refusal to publish. For example, consider two small trials both containing 30 participants. One trial shows a large positive effect size of 0.75, which is statistically significant, whilst the other shows an effect size of 0.30, which is not significant. The first trial might be accepted for publication on the basis that it has shown a large and potentially relevant benefit, whilst the other might be rejected because its sample size is too small. Because both trials have tiny sample sizes their point estimate of effect is likely to be in error. Let us assume the 'true' effect lies somewhere in between: this can be estimated through using a meta-analysis. If we combine these two small trials in a meta-analysis we can show that the 'average' effect size is 0.50, but this effect is not quite statistically significant (95% confidence interval = -0.01 to 1.02, p = 0.055). Therefore, the trial that was accepted for publication 'overestimates' the true effect, whilst the second trial that underestimates the true effect remains unpublished. On the basis of the two trials we might, therefore, call for another large and well-conducted study to confirm the suggestion of a benefit.

If publication bias were particularly severe we might identify a dozen small trials all producing 'over-estimated' effect sizes. Performing a meta-analysis of these might lead us to conclude that the large benefit justifies the cost of implementing the intervention, whereas the actual effect of

the intervention could be so small as not to justify implementation.

Because publication bias can produce misleading results it is important that its presence is detected and discussed in the review. One relatively simple way of looking for publication bias is through the use of a 'funnel plot'.

Funnel plot

A funnel plot graphically displays the effect sizes from identified trials along with some estimate of their sampling error (e.g. sample size). All trials only produce an *estimate* of the effect of the intervention, which is bounded by uncertainty. The effect of chance underpins the design and interpretation of trials. A small trial can produce some surprisingly good or poor results, merely by chance. The larger the trial the less likely is the effect of chance on the outcome. Combining small trials that have positive and negative findings has a similar effect to undertaking a single large trial, and these chance effects for positive and negative findings will balance each other out.

We can use the increased variability of small trials compared with large trials to establish whether or not there is evidence for publication bias. As the larger trials produce effect sizes closest to the 'true' value compared with small trials we can show the relationship between size and effect in a funnel plot. In a funnel plot the effect size of a trial is plotted on the x-axis against its sample size on the y. The smaller and less precise trials will be scattered along the x-axis whilst the larger and more precise studies will be clustered together. Where there is no publication bias the trials will form an inverted funnel shape, hence, the term 'funnel plot'. In Figure 6.1 a hypothetical funnel plot is shown where there is no evidence of publication bias.

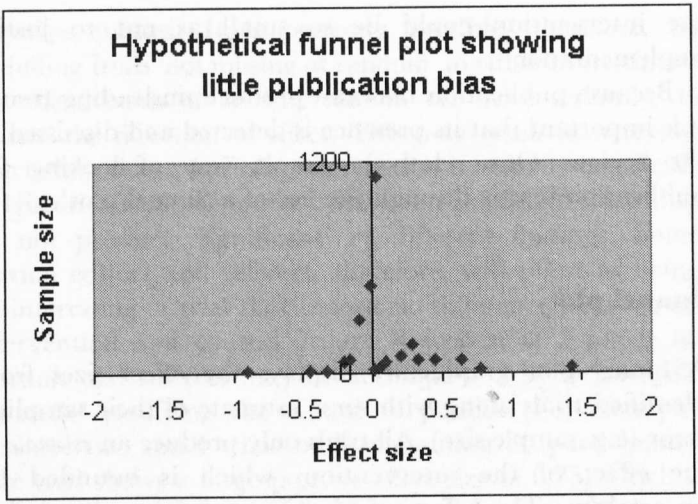


Figure 6.1 Funnel plot showing no evidence for publication bias

The figure shows a hypothetical review of a subject area where there is little evidence for any difference between the groups in terms of an overall benefit. Note how the small trials show a large variation in effect with some showing a large positive effect size of 1 or more but with others showing a similarly large negative effect size. If all of these trials were combined in a meta-analysis it is likely that it would show no overall effect.

In contrast, Figure 6.2 shows a funnel plot, taken from a recent systematic review, where there is evidence of publication bias (Torgerson, Porthouse, Brooks 2003). Note that all the trials, including the very small ones, demonstrate a positive effect. It is very unlikely, given the tiny sample sizes, that all eight trials would, by chance, have shown a positive and mostly quite large effect size. Therefore, it is likely that there are other trials that have either not been published or are only available as obscure reports, and which were not identified by the search strategy. These other trials would have either negative or null effects.

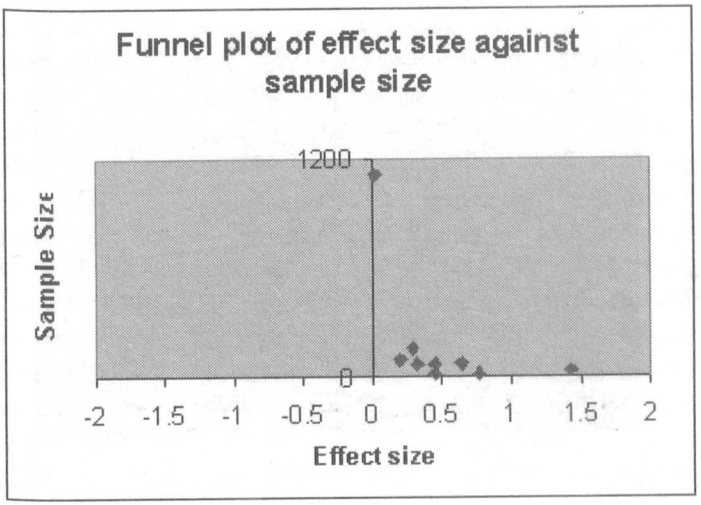


Figure 6.2 Funnel plot of RCTs in adult literacy showing publication bias

In Figure 6.3 another funnel plot shows evidence of publication bias. The data to construct the plot were taken from a systematic review by Ehri, Nunes, Willows and colleagues (2001). In the Ehri review of systematic phonics instruction interventions, one of the inclusion criteria was journal articles that had been peer-refereed. Including this criterion will potentially increase the risk of overestimating the effect size of the intervention, as it is more likely that negative studies will have been excluded. As Figure 6.3 shows, there were no studies reporting a negative effect of systematic phonics instruction compared with all forms of control despite the small sample sizes of the included studies.

Therefore, any results of a meta-analysis from trials present in the funnel plot should be treated with a high degree of caution, as they are likely to overestimate the effectiveness of the interventions.

A problem with funnel plots is that they become more unreliable at detecting the existence of publication bias in

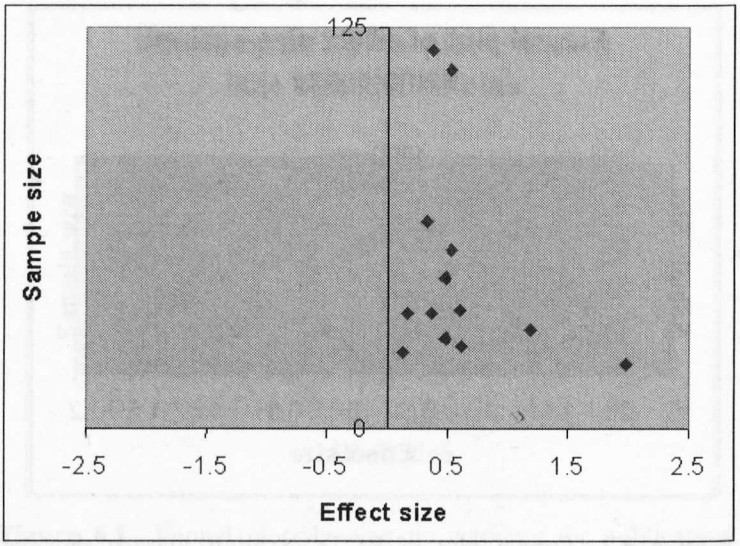


Figure 6.3 Funnel plot of randomized trials from a systematic review of systematic phonics instruction showing presence of publication bias

the presence of very few trials. For example, if in the review of studies in Figure 6.2 the two largest trials had not been undertaken or had been missed on the search strategy, the funnel plot would indicate little publication bias.

Another way of examining publication bias is to compare the effect sizes of reports published in peer review journals with those in unpublished reports or theses. Because authors, referees and journal editors are less likely to write up and publish negative trials, it is likely that trials that have been written up as a report to the funding body of the research or as a thesis for the requirements of a degree will demonstrate different effect sizes, if publication bias exists. Again, in the field of adult literacy education this appears to be the case. When the mean effect sizes of the studies that have been published (that is, formally, for example in refereed journals) are compared with those that are 'unpublished' (that is, published informally, for example

as in-house reports or mimeographs, or in the so-called 'grey' literature) we may observe a different estimate. The six published studies show a pooled effect size of 0.49 (95% CI 0.17 to 0.82, p = 0.003), whereas the three 'unpublished' studies show a lower effect size of 0.26 (95% CI −0.07 to 0.59, p = 0.13). Clearly, the unpublished studies are 'published' in the sense that their reports are obtainable and are in the public domain. Nevertheless, these reports tend to show an effect size approximately half the size of the effect sizes of the studies that are published in journals. Studies that have never been published in any form are likely to have even smaller effect sizes. These data coupled with the funnel plot indicate that there are probably significant numbers of 'missing' studies with either negative, null or very small positive effect sizes. Again this kind of result indicates that the results should be treated cautiously.

If publication bias is detected, what should be done? There are a number of methods for attempting to 'weight' the results of the review to take into account publication bias, none of which is completely satisfactory. One method, 'trim and fill' is basically to remove 'outlying' trials from the funnel plot until we have a symmetrical plot and then reintroduce the trials with a corresponding hypothetical trial with the opposite effect (Sterne *et al.* 2001). This approach, assumes however, that the missing studies have the inverse values of the outlying identified trials, which may not be the case. Another approach is to only undertake a meta-analysis on trials with large sample sizes (e.g. > 100). The reasoning behind this is that usually only small trials tend not to be published and that these missing small trials will all be negative, which may not be a realistic assumption. This approach will, however, negate an important rationale of meta-analysis: combining small studies to yield a more precise estimate of effect (Lipsey and Wilson 2001). Indeed, in the author's meta-analysis of RCTs evaluating the effectiveness of information and

communication technology on spelling, this approach would have resulted in the rejection of all of the identified studies (Torgerson and Elbourne 2002).

Sometimes studies are published with missing data. If no effect is found, authors may report that 'there were no significant differences between the groups' but may not report the mean values to allow their inclusion in a meta-analysis. Writing directly to the authors may yield the necessary data in some instances. Published studies with missing data can be included in a meta-analysis with simulated data to show no effect, which may be better than not including them at all (Lipsey and Wilson 2001).

Publication bias is a real threat to the validity of any systematic review. It is important that the reviewer should examine the data for evidence of such bias. It is also important to note that absence of evidence for such bias does not necessarily mean absence of bias. If there seems to be bias this should be highlighted in the discussion of the review and some steps can be taken to examine its influence, such as the use of sensitivity analysis, which is discussed later. It is important that all randomized controlled trials (even small underpowered trials) should be published whatever their results. Given the wide availability of electronic or web-based publishing it will be less likely in the future that articles will disappear without trace in obscure paper journals.

Quality assessment of systematic reviews

Systematic reviews can, clearly, vary in their quality. As with quality assessment of randomized trials, there have been a number of attempts to derive quality assurance scales in order to quality appraise systematic reviews (Shea *et al.* 2001). Recently a quality checklist of systematic reviews – the Quality of Reporting of Meta-Analyses (QUORUM) –

checklist has been developed, which has been compared against a number of other quality assurance scales (Shea *et al.* 2001). It may be helpful before commencing a systematic review to check that the review broadly follows the QUORUM checklist, which seems to be more comprehensive than other quality assurance scales.

Box 6.1 Key features of QUORUM statement (adapted for educational studies)

Introduction:	Explicitly state educational problem and rationale for review.
Methods	
Searching:	State sources of information (e.g., names of databases; hand searching of key journals), search restrictions (e.g., year, publication language, published and or unpublished).
Selection:	Inclusion and exclusion criteria.
Validity assessment:	Quality assessment (e.g., blinded follow-up).
Data abstraction:	Process used (e.g., double data extraction).
Study characteristics:	Type of study design, student characteristics, details of intervention, outcomes, how was educational heterogeneity assessed?
Data synthesis:	How were data combined? Measures of effect, statistical testing and confidence intervals, handling of missing data, sensitivity and subgroup analyses, assessment of publication bias.
Results	
Trial flow:	Provide a profile of trials identified and reasons for inclusion/exclusion.
Study characteristics:	Provide descriptive data for each trial (e.g., age, setting, class size, intervention).
Quantitative data synthesis:	Report agreement between reviewers on selection and validity assessment; present summary results; report data needed to calculate effect sizes and confidence intervals

	(i.e., number; mean; standard deviations by group).
Discussion:	Summarize key findings and educational inferences. Interpret results in light of all the evidence; acknowledge potential biases in review and suggest areas for future research.

Source: Shea *et al.* (2001)

Summary

- Publication bias can occur in a systematic review if studies 'not missing at random' are excluded: this can overestimate the effect or reverse the direction of effect.
- Publication bias can threaten the validity of a systematic review. Therefore it is important that its presence is detected and discussed.
- One relatively simple way of looking for publication bias is through the use of a 'funnel plot'.
- As with quality assessment of randomized trials, there have been a number of attempts to derive quality assurance scales in order to quality appraise systematic reviews.
- Before undertaking a review it may be helpful to check that it broadly follows the QUORUM checklist.

7

Data Synthesis and Meta-analysis

Data from randomized trials can be synthesized in a number of different ways. Firstly, a 'qualitative' overview of the studies can be undertaken. The basic characteristics of the studies are described, including their methodological strengths and weaknesses. This aspect of the review requires a good understanding of the methodology of trial design and execution. Subject specialism is also important when understanding the intervention. For example, a study might compare an intervention delivered through information and communication technology with the same intervention delivered through 'conventional' teaching. However, it is possible that the ICT intervention is quite different from the 'treatment' delivered in the control condition. Therefore, in such circumstances one could not disentangle the effects of ICT from the effects of the different teaching strategies.

Another way of synthesizing the identified studies is through the use of 'vote' counting (see Davies 2000 for a full discussion). For example, if ten trials were identified in a review, this method would state that six showed a positive effect (three of which were statistically significant), three showed a negative effect and one showed no effect. Vote counting may be useful in describing the overall effects of the trials, especially when a meta-analysis is not possible. However, this method should be treated with caution. If, for example, one of the ten trials that showed no effect

contained 1000 participants and was a rigorously designed and executed RCT, this study would carry more weight than the nine remaining trials (particularly if they were small and poorly designed and conducted). On the other hand, if all the trials were of similar quality and size, and there was no evidence of publication bias, vote counting may give some indication as to whether or not there could be an overall effect.

Table 7.1 is an example of a summary data synthesis table for a systematic review of the effect of unpaid classroom assistants on children's reading. It presents information about the aims of each RCT, the setting, the participants, the intervention and control 'treatments' and the outcome measurements.

Table 7.2 shows a 'qualitative' synthesis of RCTs identified in a systematic review of the effect of unpaid 'volunteers' on children's reading. Four RCTs indicated a positive effect of the intervention on outcomes, one of which was statistically significant. Three RCTs indicated a negative effect, and one was equivocal. These data suggest no evidence of a benefit of volunteer classroom assistants on reading outcomes. This finding was supported by a meta-analysis of the four most homogeneous trials (see Figure 7.1, p. 84). This indicated a small, pooled effect size of 0.19, which was not statistically significant (95% confidence interval −0.31 to 0.68, $p = 0.54$). The descriptive tables, however, show that one trial was of reasonably high quality and appeared to show consistent positive effects (Baker 2000). For the researcher proposing to undertake another trial in this area, replication of a study similar to Baker's could be worthwhile.

Meta-analysis is a statistical technique of 'pooling' data from two or more randomized trials. The value of a meta-analysis lies in the fact that it reduces the random errors experienced by a single study and can lead to a more precise estimate of the overall effect. There are a number of

Table 7.1 Description of randomized controlled trials

Study reference	Study setting and design	Intervention	Type of volunteer and training	Control	Main outcome measures
Baker *et al.* 2000	USA, individually randomized trial, teachers selected children with below average reading skills from 24 first-grade classrooms. 127 children originally randomized, a third dropped out leaving 84.	Children received 1 hour to 1 hour 30 minutes tutoring two times a week for two years (average 73 sessions, SD 10.9).	Middle-aged volunteers (gender not stated) recruited mainly from the business community given 1–2 hours of training.	Normal classroom instruction.	Woodcock Reading Mastery Test–Revised (Word Identification subtest), Oral Reading Fluency, Word and Passage Comprehension.
Elliott *et* al. 2000	UK, cluster randomized trial of intact classes from primary schools in poor socio-economical areas in north-east of England. Two parallel reception classes from each of three primary schools randomized. 140 children at start of trial, 41 dropped out.	Volunteers worked with children alongside classroom teacher all the time except for practical lessons (e.g. physical education).	Mainly mature women.	Children in control classes received normal classroom lessons from teacher.	Wechsler Objective Reading Development Scales (WORD).

Study reference	Study setting and design	Intervention	Type of volunteer and training	Control	Main outcome measures
Loenen 1989	UK, children needing help identified by staff in 16 primary schools in inner London. Children were generally poor readers. 81 were randomized and 81 completed reading comprehension and accuracy post-tests.	Two out of class 30 minute 1 to 1 sessions a week over two terms.	34 volunteers over 35 years, 28 were women, all but one were experienced volunteers. Trained with 3 × 1/5 hour sessions mainly in reading for meaning techniques.	Normal classroom teaching.	Salford Sentence Reading Test (SSRT), Primary Reading Test (PRT).
Morris 1990	USA, second- and third-grade children with low reading ability from schools in low socio-economic areas. 60 children randomized no attrition.	Children given 1 hour tutoring two times a week from 3 p.m. to 4 p.m., 50 hours over a year.	Mixed age, ranging from college students to retired people.	Normal school lessons.	Word recognition, spelling, reading aloud from passages.

Study reference	Study setting and design	Intervention	Type of volunteer and training	Control	Main outcome measures
Rimm-Kaufman 1999	USA, first-grade children (5.5 to 7 years), teachers identified children needing help in literacy. Children paired and individually randomized. 42 children randomized, no attrition.	Three times a week for 45 minutes 1 to 1 tutoring from October to May. Used phonics and reading for meaning.	Well-trained volunteers all over 60 years, half retired teachers, received extensive training for five sessions before start, two bi-monthly sessions during study.	Normal classroom (teachers unaware of who were the control children).	Letters, words, print concepts, writing, dictation and reading level.
Weiss 1988	USA, 'mildly handicapped' students from grades 3 to 6 selected by teachers as those who would benefit most from additional tuition. 17 randomized, 1 lost from control group.	1 to 1 tuition for 20–30 minutes a day for four days a week over 11 weeks with a minimum of 36 sessions (maximum 44).	12 volunteers (8 senior citizens, 9 females, 4 ex-teachers). 5 hour 1 day training plus a follow-up session three weeks later. Trained in paired reading techniques, flash cards, cloze procedure.	Normal classroom teaching.	BASIS, Basic Achievement Skills Individual Screener, Curriculum Based Measurement (Holt reading series).

Study reference	Study setting and design	Intervention	Type of volunteer and training	Control	Main outcome measures
Lee 1980	USA, third- to sixth-grade pupils from either low income groups or from minority groups, children matched in pairs and then randomized. 70 randomized, no attrition.	Pupils tutored as an after-school activity in small groups, 2 to 1 for about 2 hours two times a week.	20 college students (15 women). Undertook seven training modules over 8 weeks.	Normal classroom teaching.	Reading grade equivalent scores.

Source: Torgerson *et al.* (2002) 'Do volunteers in schools help children learn to read? A systematic review of randomised controlled trials', *Educational Studies*, 28(4), 2002.

Table 7.2 Summary of results of randomized controlled trials

Study reference	Mean effect on reading outcome measures		Standardized Difference (95% confidence interval)	Favours volunteers
	Intervention (sample size)	Control (sample size)		
Baker *et al.*				
First Grade	(n=43)	(n=41)		
Word Identification	409.2 (29.7)	398.9 (24.4)	0.38 (−0.05 to 0.81)	Y*
Oral Reading	27.8 (22.8)	18.7 (17.3)	0.45 (0.01 to 0.88)	Y
Passage Comprehension	449.3 (24.4)	443.2 (14.2)	0.30 (−0.13 to 0.73)	Y*
Second Grade				
Word Identification	449.4 (30.2)	437.9 (25.9)	0.41 (−0.03 to 0.84)	Y*
Oral Reading 1	71.3 (35.2)	55.9 (32.1)	0.36 (−0.07 to 0.79)	Y*
Oral Reading 2	61.5 (35.5)	45.9 (29.5)	0.48 (0.04 to 0.91)	Y
Word Comprehension	472.3 (17.3)	465.4 (16.2)	0.41 (−0.02 to 0.84)	Y*
Passage Comprehension	468.9 (16.0)	464.7 (13.1)	0.27 (−0.16 to 0.70)	Y*
Elliott *et al.*	(n=50)	(n=49)		
Reading Accuracy	89.8 (15.6)	90.6 (16.4)	−0.05 (−0.44 to 0.34)	N*
Reading Comprehension	88.5 (14.7)	89.6 (13.8)	−0.08 (−0.47 to 0.32)	N*
Spelling	91.7 (14.3)	93.5 (11.7)	−0.14 (−0.53 to 0.26)	N*
Composite Score	87.9 (17.0)	89.4 (15.4)	−0.09 (−0.49 to 0.30)	N*

Study reference	Mean effect on reading outcome measures		Standardized Difference (95% confidence interval)	Favours volunteers
	Intervention (sample size)	Control (sample size)		
Loenen	(n=43)	(n=38)		
Reading Comprehension	19.51 (7.68)	22.31 (7.83)	−0.36 (−0.80 to 0.08)	N*
Reading Accuracy	92.58 (10.78)	97.37 (14.247)	−0.38 (−0.82 to 0.06)	N*
Lee+	(n=20)	(n=20)		
Reading Grade Score Change	0.9675 (0.587)	0.9300 (0.759)	0.06 (-0.57 to 0.67)	Y*
Morris	(n=30)	(n=30)		
Time Word recognition	58.3 (17.4)	49.9 (19.4)	0.46 (−0.06 to 0.97)	Y*
Untimed Word recognition	77.6 (19.1)	69.8 (22.6)	0.37 (−0.14 to 0.88)	Y*
Basal Word recognition	21.3 (5.2)	18.0 (6.5)	0.56 (0.04 to 1.07)	Y
Passage Comprehension	15.3 (9.4)	9.9 (5.8)	0.69 (0.17 to 1.21)	Y
Spelling (score)	7.7 (4.2)	5.8 (4.3)	0.45 (-0.07 to 0.96)	Y*
Spelling qualitative	85.1 (18.0)	75.4 (21.1)	0.49 (-0.02 to 1.01)	Y*

Study reference	Mean effect on reading outcome measures		Standardized Difference (95% confidence interval)	Favours volunteers
	Intervention (sample size)	Control (sample size)		
Rimm-Kaufman	(n=21)	(n=21)		
Letters	52.86 (1.28)	51.62 (2.80)	0.57 (−0.05 to 1.18)	Y*
Words	13.14 (6.10)	13.38(5.64)	−0.04 (−0.65 to 0.56)	N*
Print Concepts	14.29 (3.18)	14.19 (2.82)	0.03 (−0.57 to 0.64)	N*
Writing	23.05 (11.02)	23.19 (11.40)	−0.01 (−0.62 to 0.59)	N*
Dictation	28.14 (8.43)	26.62 (8.13)	0.18 (−0.42 to 0.79)	Y*
Reading Level	5.86 (3.76)	4.43 (2.82)	0.43 (−0.18 to 1.04)	Y*
Weiss	(n=9)	(n=7)		
BASIS	41.9 (10.1)	43.2 (10.5)	−0.13 (−1.11 to 0.86)	N*
CBM	93.2 (35.2)	106.8 (56.5)	−0.30 (−1.29 to 0.70)	N*

+ Data analysed by tutor to account for clustered nature of data

* Not statistically significant

commercially available software packages that can perform a meta-analysis.

Before a meta-analysis is undertaken a decision needs to be made about whether or not the trials are educationally homogeneous. Lack of homogeneity can sometimes be obvious, such as differences between trials in adult versus child learners. Other sources of 'heterogeneity' may be less obvious, except to the content specialist. Some interventions, whilst appearing to be superficially similar, may have completely different psychological underpinnings and delivery mechanisms. If heterogeneous trials are pooled then the resulting point estimate will not apply to any of the interventions. In the meta-analysis that was undertaken of randomized trials evaluating volunteers, homogeneity was assessed in two ways: the amount of volunteer training, and learner characteristics of the participants. Two trials, therefore, were excluded from the meta-analysis: one because it used volunteers with little training, and the other because it included children who experienced learning disabilities. The reasoning behind such exclusions were as follows. Volunteering is likely to be most effective if volunteers receive training. Including a trial in a meta-analysis of 'untrained' volunteers is likely to dilute any intervention effect and introduce a source of heterogeneity. Similarly, volunteering may have different effects in relation to whether or not the children experience disabilities in learning. Therefore it would not seem sensible to meta-analyse trials including participants with different learner characteristics.

Assuming two or more trials have been identified, therefore, and have used similar interventions in similar contexts, what is the process? Most educational meta-analyses will involve calculating a pooled effect size. There are a number of different statistical approaches to meta-analysis, which are described in more detail within statistical texts specifically devoted to aspects of meta-

analysis (e.g. Lipsey and Wilson 2001). In this chapter only an overview of the main approach is described.

As described previously, an effect size is the difference in means between the two groups divided by either the pooled standard deviation or the standard deviation of the control group, to give a common metric that can be applied to studies using different forms of post-test. The standard deviation is a measure of dispersion for a continuous variable like a test score. A large value for the standard deviation indicates that there is a large spread of values around the mean value. The method for deriving the standard deviation is available from all basic statistical textbooks, and is routinely calculated for statistical output from software packages.

The advantage of calculating a standardized effect size is that it enables comparisons to be made between studies that use very different measures of outcome. For example, two spelling trials were undertaken by MacArthur *et al.* (1990) and McClurg and Kasakow (1989). The study by MacArthur *et al.* used a spelling test out of 20 for the outcome measure, whilst the McClurg and Kasakow trial used a spelling test out of 36. The difference between the groups in the McClurg and Kasakow study was an average of six spellings compared with only two spellings in the McArthur *et al.* trial. However, these differences are not directly comparable as the spelling tests were quite different. By calculating a standardized effect size Torgerson and Elbourne (2002) could show that the MacArthur trial had an effect size of 0.35, whilst the McClurg and Kasakow study had an effect size of 1.15.

Once the effect sizes of each study, with their associated 95% confidence intervals, have been calculated, the next step is to pool all the data in a meta-analysis. We cannot, however, simply average the standardized effect sizes and generate an average as this gives equal weight to all the trials, when in fact the trials with the bigger sample sizes

should be given most weight as their results are closer to the 'true' value. Therefore, in the meta-analysis the effect sizes from each trial is 'weighted' by the trial's size. Larger trials receive a greater weight to reflect their greater importance.

Usually the software program that produces the meta-analysis will also produce a graphical display of the effect size of each individual randomized trial with 95% confidence intervals. This is known as a 'forest plot'. A forest plot is a helpful graphical aid when examining all the effect sizes of the identified trials. It can be used to describe all the identified studies even when there is no intention to pool or meta-analyse them.

In Figure 7.1 a typical forest plot shows a meta-analysis of four trials evaluating the effects of using unpaid classroom assistants (volunteers) to help children learn to read. The effect size of each trial is calculated with the appropriate

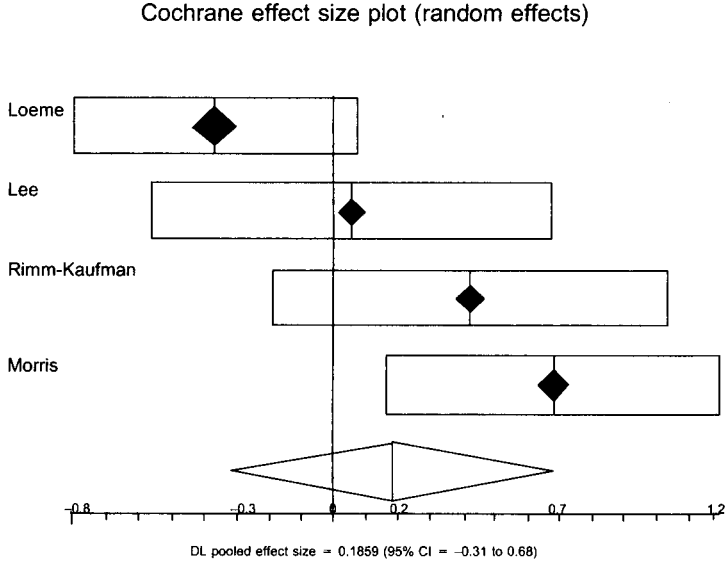


Figure 7.1 Forest plot of randomized trials of the effect of unpaid volunteers on literacy outcomes

95% confidence intervals. Small trials have very wide confidence intervals. This particular meta-analysis is suggestive of an overall benefit of about a fifth of an effect size. The 'pooled' confidence interval, however, is wide (reflecting the small trials that were included), and the overall effect size is not statistically significant. In this instance the use of volunteer assistants is suggestive of benefit but is not conclusive. Indeed, because the confidence interval passes through zero, the use of volunteers could actually worsen educational outcomes. Therefore, this meta-analysis is a powerful pointer towards the need for a large well-conducted trial of volunteer helpers in schools.

If trials use cluster or group randomization (that is the class or school is the unit of randomization) then these cannot be included in a meta-analysis with individually randomized trials.

Meaning of effect sizes

An effect size is a statistical device that facilitates a comparison of the effectiveness of different studies that use disparate measures of effect. It is important, however, to consider what an effect size means in educational practice. In Table 7.3 different effect sizes are translated into proportions passing a test.

If we apply some of these effect sizes at a population level we can see how important even quite small effect sizes can be. If for example, class sizes were reduced from 25 pupils to 15 this would result in an effect size of about 0.15. Applying this to a national school population of say 700,000 taking a particular public exam, this would imply that about 40,000 more children would cross the 50 per cent threshold if smaller class sizes were implemented. Small effect sizes can be worthwhile; however, to detect these modest but important differences requires either a very large trial, like

Table 7.3 Translating effect sizes into pass rates

Mean effect size	Percentage of extra students passing a 50% test threshold
0	0
0.1	4
0.2	8
0.3	12
0.4	16
0.5	19
0.6	23
0.7	26
0.8	29
0.9	32
1.0	34

the Tennessee trial of class sizes, or a systematic review combined with a meta-analysis.

Sensitivity analysis

Sensitivity analyses can be undertaken to test the robustness of the review results. One way of testing the results is to undertake separate analyses on subgroups of trials. For example, do trials that use blinding or masking of post-tests generate similar effect sizes to trials that do not state that blinding was used? Are the results from large trials different from the results from small trials?

Summary

- Data from randomized trials can be synthesized in a number of different ways: 'qualitative' overviews, 'vote-

counting' methods, meta-analyses, all of which have their strengths and limitations.

- Meta-analysis allows the detection of small and possibly educationally important effect sizes.
- Whether or not small effect sizes are 'worthwhile' depends upon the nature of the intervention and its cost.

8

Conclusions

Systematic reviews of randomized controlled trials are very important policy tools. If reviews show little evidence of an effect this is an extremely important result. Large amounts of resources are often directed toward interventions where there is little or no evidence of effectiveness. For example, non-randomized data from both the UK and Israel indicate that the large investment in ICT in education may be counter-productive. Israeli data suggest ICT is actually harmful in the learning of mathematics and has no effect in the learning of Hebrew. Systematic reviews of randomized trials of ICT and literacy from the English-speaking world show little evidence of benefit of ICT on the acquisition of literacy, supporting the worrying findings from the non-randomized data (Torgerson and Zhu 2003).

Negative reviews also sound cautionary notes about the need to base policy-making on robust trial data. Whilst undertaking a large rigorous trial of the effectiveness of ICT on literacy would be a relatively expensive piece of research, the cost pales into insignificance compared with the cost of not doing the research.

Similarly, a systematic review of programmes for adult literacy and numeracy uncovered little evidence for any single effective method of improving literacy and numeracy skills among adults (Torgerson, Porthouse, Brooks 2003). Again such a finding is extremely important in terms of the needs for research.

On a positive note, systematic reviews have been

instrumental in showing smaller class sizes are related to improved performance in educational outcomes. Also systematic reviews have shown that phonological awareness training and phonics teaching are effective for improving literacy acquisition among young children (Ehri, Nunes, Stahl, Willows 2001, Ehri, Nunes, Willows *et al.* 2001).

For primary researchers meta-analyses and systematic reviews are important. Ideally, a positive finding of an intervention that is observed in a meta-analysis ought to be confirmed by an appropriately designed trial. Whilst a 'follow-up' trial *can* confirm the findings of a meta-analysis (Hedges 2000), this is not always the case (Fukkink 2002).

Although systematic reviews ought to be an important tool for the policy-maker, even very persuasive findings are not always implemented. The UK government insists on implementing driver education despite strong evidence showing its lack of effectiveness. Nevertheless, one of the aims of research is to reduce uncertainty, and whilst some policy-makers may wish to operate in an evidence-free environment this will not always be the case.

This book has described the first steps that will enable a student or researcher to undertake a systematic review. An important aspect of systematic reviewing, which perhaps has not been emphasized sufficiently, is its collaborative nature. Whilst a systematic review can be undertaken by a single reviewer, and described as such, the quality of the review process is undoubtedly improved by collaboration with various specialists, including information and content specialists, trial methodologists and statisticians.

Systematic reviews will inevitably be 'good, bad and indifferent'. What sets a systematic review apart from the other research tool – the narrative review – is that a systematic review can be re-examined because its methods are explicit and replicable. Controversial findings from a systematic review can be tested using either the same criteria as described by the authors to check for errors or,

alternatively, different criteria can be used to include, exclude or combine studies. Often valid criticisms of systematic reviews can be made because they *are* so explicit. This is in contrast to non-systematic reviews, where the methods can often be opaque.

Systematic review techniques were pioneered by educational researchers, but have, to an extent, fallen out of fashion. There is a welcome increased interest in the technique by newer generations of educational researchers.

Suggested Further Reading

Chalmers, I. (2001) 'Comparing like with like: some historical milestones in the evolution of methods to create unbiased comparison groups in therapeutic experiments', *International Journal of Epidemiology*, 30, 1156–64.

Chalmers, I., Hedges, L. V. and Cooper, H. (2002) 'A brief history of research synthesis', *Evaluation and the Health Professions*, 25, 12–37.

Cook, T. D. (2002) 'Reappraising the arguments against randomized experiments in education: an analysis of the culture of evaluation in American schools of education.' Paper for presentation at the SRI International Design Conference on Design Issues in Evaluating Educational Technologies (www.sri.com/policy/designkt/found.html).

Oakley, A. (2000) *Experiments in Knowing: Gender and Method in the Social Sciences*. Cambridge: Polity Press.

Slavin, R. E. (1986) 'Best-evidence synthesis: an alternative to meta-analytic and traditional reviews', *Educational Researcher*, 15, 5–11.

References

Altman, D. G. (1996) 'Better reporting of randomised controlled trials: The CONSORT statement', *British Medical Journal*, 313, 570–1.

Badger, D., Nursten, J., Williams, P. and Woodward, M. (2000) 'Should all literature reviews be systematic?', *Evaluation and Research in Education*, 14, 220–30.

Batchelder, J. S. and Rachal, J. R. (2000) 'Effects of a computer-assisted instruction programme in a prison setting: an experimental study', *Journal of Correctional Education*, 51, 324–32.

Berlin, J. A. (1997) 'Does blinding of readers affect the results of meta-analyses?', *Lancet*, 350, 185–6.

Boruch, R. F. (1994), 'The future of controlled randomised experiments: a briefing', *Evaluation Practice*, 15, 265–74.

Boruch, R. F. (1997) 'Randomized experiments for planning and evaluation: a practical approach', in *Applied Social Research Methods Series 44*. London: Age Publications.

Chalmers, I. (2001) 'Comparing like with like: some historical milestones in the evolution of methods to create unbiased comparison groups in therapeutic experiments', *International Journal of Epidemiology*, 30, 1156–64.

Chalmers, I. and Altman, D. G. (eds) (1995) *Systematic Reviews*. London: BMJ Publishing Group.

Chalmers, I., Hedges, L. V. and Cooper, H. (2002), 'A brief history of research synthesis', *Evaluation and the Health Professions*, 25, 12–37.

Clayton, A. B., Lee, C., Sudlow, D. E., Butler, G. and Lee, T. (1998) *Evaluation of the DSA school's initiatives*. British Institute of Traffic Education Research.

Cochrane Injuries Group Albumin Reviewers (1998) 'Human albumin administration in critically ill patients: systematic review of randomised controlled trials', *British Medical Journal*, 317, 235–40.

Cochrane Injuries Group Driver Education Review (2001) 'Evidence based road safety: the Driving Standards Agency's school programme', *Lancet*, 358, 230–2.

Constable, H. and Coe, R. (2000) 'Evidence and indicators: dialogue, improvement and researching for others', *Evaluation and Research in Education*, 14, 115–23.

Cook, T. D. (2002) 'Reappraising the arguments against randomized experiments in education: an analysis of the culture of evaluation in American schools of education.' Paper for presentation at the SRI International Design Conference on Design Issues in Evaluating Educational Technologies (www.sri.com/policy/designkt/found.html).

Cook, T. D. and Campbell, T. D. (1979) *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin.

Crain, R. L. and York, R. L. (1976) 'Evaluating a successful program: experimental method and academic bias', *School Review*, 84, 233–54.

Davies, H., Nutley, S. and Smith, P. (2000), 'Introducing evidence-based policy and practice in public services', in H. Davies, S. Nutley and P. Smith (eds), *What Works? Evidence-based Policy and Practice in Public Services*. Bristol: The Policy Press, 1–11.

Davies, H., Nutley, S. and Tilley, N. (2000), 'Debates on the role of experimentation', in H. Davies, S. Nutley and P. Smith (eds), *What Works? Evidence-based Policy and Practice in Public Services*. Bristol: The Policy Press, 251–76.

Davies, H., Laycock, G., Nutley, S., Sebba, J. and Sheldon, T. (2000) 'A strategic approach to research and

development', in H. Davies, S. Nutley and P. Smith (eds), *What Works? Evidence-based Policy and Practice in Public Services*. Bristol: The Policy Press, 229–50.

Davies, P. (1999) 'What is evidence-based education?', *British Journal of Educational Studies*, 47, 108–21.

Davies, P. (2000) 'The relevance of systematic reviews to educational policy and practice', *Oxford Review of Education*, 26, 365–78.

DfEE (1989) *The National Literacy Strategy. Framework for Teaching*: London: DfEE.

Egger, M. and Davey-Smith, G. (2001) 'Principles of and procedures for systematic reviews', in M. Egger, G. Smith and D. Altman (eds), *Systematic Reviews in Health Care Meta-Analysis in Context* (2nd edn). London: BMJ Books.

Egger, M., Davey-Smith, G. and O'Rourke, K. (1995) 'Rationale, potentials and promise of systematic reviews', in I. Chalmers and D. Altman (eds), *Systematic Reviews*. London: BMJ Publications Group.

Egger, M., Smith, G. D. and Altman, D. G. (eds) (2001) *Systematic Reviews in Health Care Meta-Analysis in Context* (2nd edn). London: BMJ Books.

Ehri, L. C., Nunes, S. R., Stahl, S. A. and Willows, D. M. (2001) 'Systematic phonics instruction helps students learn to read: evidence from the national reading panel's meta-analysis', *Review of Educational Research*, 71, 393–447.

Ehri, L. C., Nunes, S. R., Willows, D. M., Valeska Schuster, B., Yaghoub-Zadeh, Z. and Shanahan, T. (2001) 'Phonemic awareness instruction helps children learn to read: evidence from the National Reading Panel's meta-analysis', *Reading Research Quarterly*, 36 (3), 250–87.

Evans, J. and Benefield, P. (2001) 'Systematic reviews of educational research: does the medical model fit?', *British Educational Research Journal*, 27 (5), 528–41.

Eysenck, H. J. (1995) 'Problems with meta-analysis', in I. Chalmers and D. Altman (eds), *Systematic Reviews*. London: BMJ Publication Group.

Fitz-Gibbon, C. (2000) 'Education: realising the potential', in H. Davies, S. Nutley and P. Smith (eds), *What Works? Evidence-based Policy and Practice in Public Services*. Bristol: The Policy Press, pp. 69–91.

Fukkink, R. G. (2002) 'Effects of instruction on deriving word meaning from context and incidental word learning', *Educational Studies in Language and Literature*, 2, 38–57.

Fulk, B. M. and Stormont-Spurgin, M. (1995) 'Spelling interventions for students with disabilities: a review', *The Journal of Special Education*, 28 (4), 488–513.

Glass, G. V., (1976) 'Primary, secondary and meta-analysis', *Educational Researcher*, 5, 3–8.

Gough, D. and Elbourne, D. (2002) 'Systematic research synthesis to inform policy, practice and democratic debate', *Social Policy in Society*, 1 (3), 225–36.

Hammersley, M. (1997) 'Educational research and a response to David Hargreaves', *British Educational Research Journal*, 23 (2), 141–61.

Hammersley, M. (2001) 'On "systematic" reviews of research literatures: a "narrative" response to Evans and Benefield', *British Educational Research Journal*, 27 (5) 543–53.

Hargreaves, D. H. (1996) 'Teaching as a Research-based Profession: Possibilities and Prospects.' Teacher Training Agency Annual Lecture. London: Teacher Training Agency.

Hargreaves, D. H. (1997) 'In defence of research for evidence-based teaching: a rejoinder to Martyn Hammersley', *British Educational Research Journal*, 23 (4), 405–19.

Hatcher, P., Hulme, C. and Ellis, A. (1994) 'Ameliorating early reading failure by integrating the teaching of reading and phonological skills: the phonological linkage hypothesis', *Child Development*, 65, 41-57.

Hedges, L. V. (2000) 'Using converging evidence in policy formation: the case of class size research', *Evaluation and Research in Education*, 14, 193–205.

Juni, P., Altman, D. G. and Egger, M. (2001) 'Assessing the quality of randomised controlled trials', in M. Egger, G. Davey-Smith and D Altman (eds), *Systematic Reviews in Health Care: Meta-analysis in Context* (2nd edn). London: BMJ Publishing Group.

Juni, P., Witschi, A., Bloch, R. and Egger, M. (1999) 'The hazards of scoring the quality of clinical trials for meta-analysis', *JAMA*, 282, 1054–60.

Kjaergard, L. L., Villumsen, J. and Cluud, C. (2001) 'Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses', *Annals of Internal Medicine*, 135, 982–9.

Kulik, J. and Kulik, C. C. (1989) 'Meta-analysis in education', *International Journal of Educational Research*, 13 (3), 220.

Lipsey, M. W. and Wilson, D. B. (1993) 'The efficacy of psychological, educational and behavioural treatment: confirmation from meta-analysis', *American Psychologist*, 48, 1181–209.

Lipsey, M. W. and Wilson, D. B. (2001) *Practical Meta-Analysis*. Applied Social Research Methods Series 49. London: Sage Publications.

MacArthur, C. A., Ferretti, R. P., Okolo, C. M. and Cavalier, A. R. (2001) 'Technology applications for students with literacy problems: a critical review', *The Elementary School Journal*, 101 (3), 273–301.

MacArthur, C. A., Haynes, J. A., Malouf, D. B., Harris, K. and Owings, M. (1990) 'Computer assisted instruction with learning disabled students: achievement, engagement and other factors that influence achievement', *Journal of Educational Computing Research*, 6, 311–28.

Martinson, K. and Friedlander, D. (1994) *GAIN: Basic Education in a Welfare-to-Work Program. California's Greater Avenues for Independence Program*. New York: Manpower Demonstration Research Corporation.

McClurg, P. A. and Kasakow, N. (1989) 'Wordprocessors, spelling checkers, and drill and practice programs: effective tools for spelling instruction?', *Journal of Educational Computing Research*, 5, 187–98.

Mulrow, C. (1994) 'Rationale for systematic reviews', *British Medical Journal*, 309, 597–9.

NHS Centre for Reviews and Dissemination (2001) *Undertaking systematic reviews of research on effectiveness: CRD's guidance for those carrying out or commissioning reviews* (CRD Report No. 4: 2nd edn). York: University of York, NHS Centre for Reviews and Dissemination.

Oakley, A. (1998) 'Experimentation and social interventions: a forgotten but important history', *British Medical Journal*, 317, 1239–42.

Oakley, A. (2000) *Experiments in Knowing: Gender and Method in the Social Sciences*. Cambridge: Polity Press.

Oakley, A. (2002) 'Social science and evidence-based everything: the case of education', *Educational Review*, 54, 277–86.

Petrosino, A., Boruch, R. F., Rouding, C., McDonald, S. and Chalmers, I. (2000) 'The Campbell Collaboration Social Science, Psychological, Educational and Criminological Trials Register (C2-SPECTR) to facilitate the preparation and maintenance of systematic reviews of social and educational interventions', *Evaluation and Research in Education*, 14, 206–19.

Petrosino, A., Turpin-Petrosino, C. and Buehler, J. (2002) '"Scared straight" and other juvenile awareness programmes for preventing juvenile delinquency.' Campbell Collaboration website www.campbellcollaboration.org/doc-pdf/ssr.pdf

Petticrew, M. (2001) 'Systematic reviews from astronomy to zoology: myths and misconceptions', *British Medical Journal*, 322, 98–101.

Pirrie, A. (2001) 'Evidence-based practice in education: the best medicine?', *British Journal of Educational Studies*, 49, 124–36.

Pring, R. (2000) 'Editorial: Educational research', *British Journal of Educational Studies*, 48, 1–10.

Roberts, I. (2000) 'Randomised trials or the test of time? The story of human albumin administration', *Evaluation and Research in Education*, 14, 231–6.

Schulz, K. F., Chalmers I., Hayes, R. and Altman, D. G. (1995) 'Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials', *JAMA*, 273, 408–12.

Shea, B., Dube, C. and Moher, D. (2001) 'Assessing the quality of reports of systematic reviews: the QUOROM statement compared to other tools', in M. Egger, G. Davey-Smith and D. G. Altman (eds), *Systematic Reviews in Health Care: Meta-analysis in Context* (2nd edn). London: BMJ Publishing Group.

Slavin, R. E. (1986) 'Best-Evidence synthesis: an alternative to meta-analytic and traditional reviews', *Educational Researcher*, 15, 5–11.

Sterne, J. A., Egger, M. and Davey-Smith, G. (2001) 'Investigating and dealing with publication and other biases', in M. Egger, G. Davey-Smith and D. G. Altman (eds), *Systematic Reviews in Health Care: Meta-analysis in Context* (2nd edn). London: BMJ Publishing Group.

Tooley, J. and Darby, D. (1998) *Education Research: An Ofsted Critique*. London: OFSTED.

Torgerson, C. J. and Elbourne, D. (2002) 'A systematic review and meta-analysis of the effectiveness of information and communication technology (ICT) on the teaching of spelling', *Journal of Research in Reading*, 35 (2), 129–43.

Torgerson, C. J. and Torgerson, D. J. (2001) 'The need for randomised controlled trials in educational research', *British Journal of Educational Studies*, 49, 316–28.

Torgerson, C. J. and Zhu, D. (2003) 'A systematic review and meta-analysis of the effectiveness of ICT on literacy

learning in English, 5–16' *(EPPI-Centre Review, version 1)*, in *Research Evidence in Education Library.* London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Torgerson, C. J., King, S. E. and Sowden, A. J. (2002) 'Do volunteers in schools help children to learn to read? A systematic review of randomised controlled trials', *Educational Studies* , 28, 433–44.

Torgerson, C. J., Porthouse, J. and Brooks, G. (2003), 'A systematic review and meta-analysis of randomised controlled trials evaluating interventions in adult literacy and numeracy', *Journal of Research in Reading*, 26 (3).

Torgerson, C. J., Brooks, G., Porthouse, J., Burton, M., Robinson, A., Wright, K. and Watt, I. (2003) *Adult literacy and numeracy interventions: Expert, scoping and systematic reviews of the trial literature.* A report to the National Research and Development Centre.

Torgerson, D. J. and Torgerson, C. J. (2003a) 'The design and conduct of randomised controlled trials in education: lessons from health care', *Oxford Review of Education*, 29, 67–80.

Torgerson, D. J. and Torgerson, C. J. (2003b) 'Avoiding bias in randomised controlled trials in educational research', *British Journal of Educational Studies*, 51 (1), 36–45.

Troia, G. A. (1999) 'Phonological awareness intervention research: a critical review of the experimental methodology', *Reading Research Quarterly*, 34, 28–52.

Ukoumunne, O. C., Gulliford, M. C., Chinn, S., Sterne, J. A. C. and Burney, P. G. J. (1998), 'Evaluation of healthcare interventions at area and organization level', in N. Black, J. Brazier, R. Fitzpatrick and B. Reeves, *Health Services Research Methods. London:* BMJ Publications.

Weiner, S. (1994) 'Effects of phonemic awareness training on low- and middle-achieving first graders' phonemic awareness and reading ability', *Journal of Reading Behavior*, 26(3), 277–300.

Wilkes, M. M. and Navickis, R. J. (2001) 'Patient survival after human albumin administration. A meta-analysis of randomized, controlled trials', *Annals of Internal Medicine*, 135, 149–64.

Young, K., Ashby, D., Boaz, A. and Grayson, L. (2002) 'Social science and the evidence-based policy movement', *Social Policy and Society*, 1, 215–24.

Index

'The Continuum Research Methods series aims to provide undergraduate, masters and research students with accessible and authoritative guides to research methodology' **RICHARD ANDREWS, SERIES EDITOR**

Every piece of primary research ought to be preceded by a systematic review. The key advantage of a systematic review over the traditional narrative review is its ability to identify all the available evidence in a systematic and replicable manner. This informative guide describes the key steps to undertaking a systematic review and includes examples of how to design data extraction forms and research strategies.

Carole Torgerson is Research Fellow in the Department of Educational Studies at the University of York.

Cover design by Helen Garvey

PRINTED IN GREAT BRITAIN
ISBN 0-8264-6580-3